CRANFIELD UNIVERSITY

Khaled Paul Taalab

Modelling Soil Bulk Density using Data-Mining and Expert Knowledge

School of Applied Sciences

Doctor of Philosophy

Supervisors: Dr. R. Corstanje, Dr. M.J. Whelan and Dr. R. Creamer

April 2013

CRANFIELD UNIVERSITY

SCHOOL OF APPLIED SCIENCES

DOCTOR OF PHILOSOPHY

Academic Year 2012 - 2013

Khaled Paul Taalab

Modelling Soil Bulk Density using Data-Mining and Expert Knowledge

Supervisors: Dr. R. Corstanje, Dr. M. Whelan and Dr. R. Creamer
April 2013

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

ABSTRACT

Data about the spatial variation of soil attributes is required to address a great number of environmental issues, such as improving water quality, flood mitigation, and determining the effects of the terrestrial carbon cycle. The need for a continuum of soils data is problematic, as it is only possible to observe soil attributes at a limited number of locations, beyond which, prediction is required. There is, however, disparity between the way in which much of the existing information about soil is recorded and the format in which the data is required. There are two primary methods of representing the variation in soil properties, as a set of distinct classes or as a continuum. The former is how the variation in soils has been recorded historically by the soil survey, whereas the latter is how soils data is typically required. One solution to this issue is to use a soil-landscape modelling approach which relates the soil to the wider landscape (including topography, land-use, geology and climatic conditions) using a statistical model.

In this study, the soil-landscape modelling approach has been applied to the prediction of soil bulk density (D_b). The original contribution to knowledge of the study is demonstrating that producing a continuous surface of D_b using a soil-landscape modelling approach is that a viable alternative to the 'classification' approach which is most frequently used. The benefit of this method is shown in relation to the prediction of soil carbon stocks, which can be predicted more accurately and with less uncertainty. The second part of this study concerns the inclusion of expert knowledge within the soil-landscape modelling approach. The statistical modelling approaches used to predict D_b are data driven, hence it is difficult to interpret the processes which the model represents. In this study, expert knowledge is used to predict D_b within a Bayesian network modelling framework, which structures knowledge in terms of probability.

This approach creates models which can be more easily interpreted and consequently facilitate knowledge discovery, it also provides a method for expert knowledge to be used as a proxy for empirical data. The contribution to knowledge of this section of the study is twofold, firstly, that Bayesian networks can be used as tools for data-mining to predict a continuous soil attribute such as D_b and that in lieu of data, expert knowledge can be used to accurately predict landscape-scale trends in the variation of D_b using a Bayesian modelling approach.

Keywords:

Bayesian networks, Random Forest, Artificial Neural Networks, Carbon Stocks, Elicitation, Soil Taxonomy, Legacy Data

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors Ron Corstanje, Mick Whelan and Rachel Creamer for their support and guidance throughout the project.

A number of people have contributed to this thesis over the years and I would like to thank Thomas Mayr, Joanna Zawadzka, Bob Jones, Jack Hannam, Gero Jahns, Pat Sills, Ian Truckell, David Parsons, Caroline Keay, Timothy Farewell, Daniel Simms and Iolanda Simo for their advice and assistance.

This project would not have been possible without the Walsh Fellowship provided by Teagasc and I would like to thank all the students and staff at Johnstown castle for making my time in Ireland a hugely enjoyable experience. In particular, I would like to thank all the soil surveyors working on the ISIS project who helped me to complete my fieldwork.

I would like to thank my friends and family for their love and support throughout my studies. Finally, I would like to thank Clare for her patience, support, encouragement and for making the last year more fun than should be possible while writing up a PhD.

TABLE OF CONTENTS

LIST OF FIGURES

5.5C.5

LIST OF TABLES

Appendix C

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AAR	Average Annual Rainfall
ANN	Artificial Neural Networks
AT0_Annual	Average Annual Accumulated Temperature Above 0°C
BN	Bayesian Network
ClORPT	Climate, Organisms, Relief, Parent Material, Time
CPT	Conditional Probability Table
D_b	Bulk Density
FCD_MED	Median Annual Number of Field Capacity Days
GIS	Geographical Information Systems
MLR	Multiple Linear Regression
OC	Organic Carbon
PSMD	Potential Soil Moisture Deficit
PT	Potential Evapotranspiration
PTF	Pedotransfer Function
RF	Random Forest
SCORPAN	Soil, Climate, Organisms, Relief, Parent Material, Age, Spatial Position
SWI	SAGA Wetness Index

1 Literature Review

This chapter reviews current methods for the prediction of soil bulk density (D_b) in relation to trends in digital soil mapping. The first section of the literature review explores the different ways in which the variation of soil and soil properties can be represented. This is an important distinction as it has both theoretical and practical implications for how predictions are made. Next the review distinguishes between maps made by the traditional soil survey and those produced using digital soil mapping techniques. Following this, there is an examination of the different statistical methods which can be used to produce soil maps. This makes particular reference to D_b, which is the property of interest in this study. The review ends by assessing the role of expert knowledge in the soil mapping process.

1.1 Introduction

1.1.1 Soil Variation

Soil properties vary almost continuously across the landscape. However, there is a limit to the number of direct observations (visual inspection, sampling and measurement) which can be made. There is, therefore, a need for prediction of soil properties at locations for which there are few or no observations (Heuvelink & Webster, 2001). The method of prediction will generally be dictated by how the soil is to be represented. Discrete units or polygons are commonly used to represent soil classes in traditional soil surveys. Discrete methods split soils into relatively homogenous groups, based on similarities in a range of properties using hierarchical classification systems (e.g. Avery, 1980). Such polygon-based classification systems are characterised by within-class homogeneity and sharp boundaries between classes. The alternative approach is to attempt to represent explicitly soils as a continuum. In this case, typically, soil attributes

(physical and morphological properties) rather than classes are mapped directly. This method is founded in geostatistics, where the values of a soil property between observations are inferred on the basis of spatial autocorrelation (Goovaerts, 1999). As this method is inextricably linked to digital information, the product of this mapping procedure is typically a grid of cells which represents the landscape, with each cell representing a predefined spatial extent, assigned a value of the property of interest. In reality, the distinction between the two approaches is not so explicit, and there are numerous examples of where the two have been integrated (Voltz & Webster, 1990; De Gruijter et al., 1997; Rawlins et al., 2008). Nevertheless, the reason that it is important to distinguish between polygon and gridded approaches is that either approach can be used to represent real-life soil variation.

1.1.2 Bulk Density (D_b)

The focus of this thesis is the spatial prediction of soil bulk density (D_b) which is defined as the oven-dry mass per unit volume of a soil (IUSS 20 Working Group, 2006). Bulk density is an important property because it is required for the calculation of the amount of soil solids present in a given area and over a given depth. It is, hence, an essential parameter in soil carbon and nutrient stock assessment (Ellert & Bettany, 1995) and in the calculation of pollutant mass balance in soil (e.g. pesticides and sewage sludge-associated synthetic organic compounds). In terms of soil classification, D_b is a key determinant of the packing structure of soils (Dexter, 1988) and is of interest for land management as it can be indicative of drainage characteristics (Arya, & Paris, 1981) such as whether there are impermeable layers in the soil, which can result in poor drainage or problems with root penetration (Lampurlanés & Cantero-Martínez, 2003). The reason that D_b is of interest in the context of mapping soil variation is that it is

seldom predicted beyond the point scale, meaning there is scope to investigate mapping D_b at the landscape scale as both a set of polygons and a continuous grid.

1.2 Representing Soil Variation

1.2.1 Polygons

The traditional method of representing spatial variation within the soil is mapping based on a soil survey. This is defined as "the process of determining the pattern of soil cover, characterising it, and presenting it in understandable and interpretable form to various users" (Rossiter, 2005). This results in a chloropleth or polygon map of soils (Figure 1-1a). Scull et al. (2003) identify three stages in the traditional survey process;

1. The direct observation of data, incorporating soil profile characteristics and ancillary data such as aerial photography, parent material and vegetation maps.
2. Incorporating observations into a conceptual model used to determine soil variation.
3. Applying the conceptual model to areas with no observed data in order to produce a predictive map.

The conceptual model used by the soil surveyors is based on understanding of the soil-landscape relationship identified by Dokuchaev (1883) (in Bockheim et al., 2005). This was further developed to represent a set of 'soil forming factors' (Jenny, 1941). These factors are given the acronym ClORPT which stands for climate (Cl), organisms (O), relief (R), parent material (P) and time (T) to illustrate the important factors influencing soil formation and variation. This was later updated to SCORPAN (McBratney et al., 2003) to include other soil properties (S), and spatial position (N) (with time (T) replaced by age (A)). Conceptually, this relies on the assumption that soils are best represented as a set of distinct classes rather than a continuum. The rational behind this

approach is that while soils do vary gradually over relatively small spatial scales, they are characterised in the wider landscape by a series of abrupt changes to a number the soil forming factors over relatively small geographic areas. This leads to 'discontinuities' in the continuum of soil and marks the boundaries between classes (Hudson, 1992). The soil-landscape approach to soil mapping, as applied in the traditional soil survey, is reliant on the soil surveyors' ability to mentally disaggregate the landscape into 'soil-landscape' units. These are generally sub-divisions of landscape features. For example, a landscape feature might be slope, while a soil-landscape unit may be a south facing slope (if the soils on slopes with other aspects were notably different). This disaggregation relates the interactions between the ClORPT variables to the soil, and is location specific. By identifying patterns in the occurrence of soil-landscape units across the landscape, it is possible to infer the soil characteristics in unsampled locations. Delineating boundaries between soil classes in this manner makes it possible to map a large number of soils using relatively few direct observations.

There are a number of issues with this approach, a particular concern is the reliance on tacit knowledge to produce the map (Zhu et al., 2001). The ability to accurately identify these soil-landscape units and the subsequent ability to relate these characteristics to a host of soil classes typically takes between 2 and 3 years experience to develop (Hudson, 1992). Once this knowledge is acquired, it is rarely recorded, meaning that much of the information regarding the landscape contained within the soil polygon map is lost, or at least hidden from those without the experience to interpret it. This means that the final product is often a map produced using unknown assumptions, leading to unknown accuracy and limitations (Scull et al., 2005). Moreover, there are problems caused by the use of the polygons themselves. Representing soil variation as a set of

polygons limits the size of the individual soil unit which can be represented. For instance, if there is a soil distinct from the surrounding soils, but covering only a small geographic space it must either be overlooked or assimilated into a surrounding soil polygon. If it is assimilated, the classification needs to be amended to include the variety within the polygon. However, there is no representation of the spatial variation of soil properties within a single polygon. Depending on the scale of the map, hundreds of hectares of soil can be assimilated into larger classes (Zhu et al., 2001). A further issue is that the polygons generalise local variation in soil properties. All soils within a polygon are considered to conform to the typical properties of the class, despite much of the soil within the polygon varying from these general attributes (Zhu, 2000). This is problematic because accurate soils data are also required as inputs for large-scale, land-surface climatic models (Best et al., 2011). This has lead to the suggestion that data collected by the traditional soil survey is inadequate for the requirements of modern users (Scull et al., 2005).

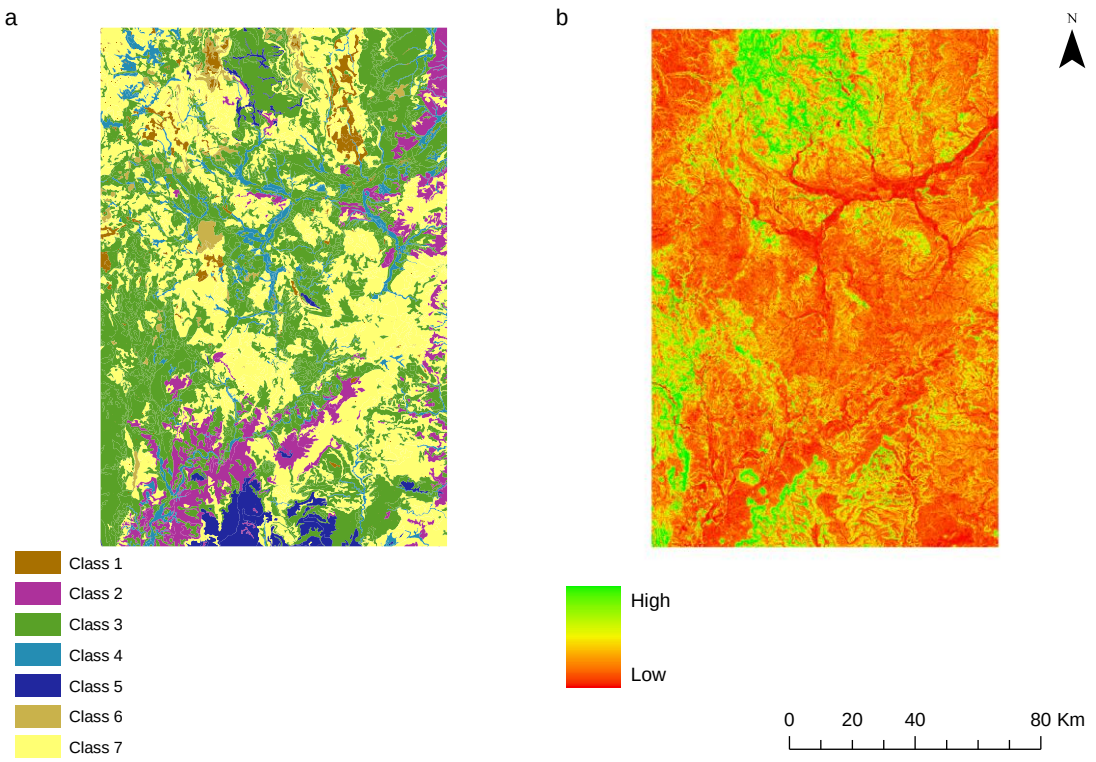


Figure 1-1: An example of the polygon and gridded approaches to mapping soil spatial variation a) Soil class represented by a set of polygons b) A single soil attribute represented on a continuous grid from high to low.

1.2.2 Gridded approaches

The representation of soil variation as a grid of cells (Figure 1-1b) rather than as distinct polygons was borne out of advances in computing technology. Webster et al. (1979) forecast that computing methods of soil mapping would change how soil survey data was processed and that with sufficient data it might be desirable to produce soil maps without distinct class boundaries. This early attempt to represent soil using gridded data used high density sampling on a regular grid to capture a range of soil properties on a relatively small scale (10 x 10km grid). Each sample was used to represent a 1 hectare grid cell and did not rely on interpolation. This approach made use of a computer to store a much larger amount of information than could be recorded in a traditional

survey. However, it was unfeasible at a larger scale due to the large number of samples required.

The ability to map soils as a grid on the same scale as traditional soil survey polygons required the application of geostatistics (Burgess & Webster, 1980). This approach adopts a fundamentally different view on the variation of soil properties across the landscape. The soil-landscape model of soil mapping (Jenny, 1941) is a deterministic method of soil mapping. Essentially, the spatial variation in soil attributes can be related to variation in the soil forming factors and, therefore, can be explained and predicted using these terms. Although there are many hundreds of different landscape variables that can be used to represent the suite of soil forming factors, it is understood that there will always be elements of these factors, or the interactions between them, which cannot be accounted for. The limitations of the models to represent real-life process are included in the error associated with prediction. Geostatistics operate under a different set of assumptions. While in essence soil formation must be deterministic, obeying certain physical laws, it can be considered random if these physical laws cannot be adequately represented. This can often be attributed to a lack of understanding regarding the processes driving spatial variations, as well as the interactions between these processes. This can be considered the case for soil, as the current state of the soil for all areas cannot be known and the interactions between the multitude of soil forming factors cannot be accurately represented (Webster, 2000).

This theoretical standpoint was borne out of the limitations of the classic soil survey approach, namely, the inability to relate the variability of certain soil attributes to soil classes and the fact that there were only two kinds of prediction possible, a class mean or a point prediction (Webster 2000). In contrast, geostatistics treat the variation within

the soil as random and interpolates variation using Kriging and other statistical methods (Goovaerts, 1999). One point to emphasise is that it is the geostatistical model which is random (stochastic), not the soil. Using geostatistics allows the prediction of a continuous surface of soil properties, under the assumption that the model residuals are spatially autocorrelated. There have been numerous variations on Kriging to improve and adapt the method. One particularly noteworthy development is 'block kriging' (Burgess & Webster, 1980) which was developed to 'scale up' measurements from soil cores in order to make predictions across large areas, avoiding the need for the gridded sampling scheme used by Webster (1979). Many more variants of Kriging are described by Cressie (1991).

There are a number of issues involving geostatistical modelling which it is necessary to be aware of. One of these relates to the assumption of stationarity within the predictive model. To illustrate this assumption, imagine that a soil sample is taken at a given location. While this sample has a definite value for any soil attribute you may wish to measure, you have to envisage that this value is simply one of an infinite number which may be there. This infinite range has a mean and variance and these do not vary across the study area (Webster, 2000). In reality this assumption probably does not hold true across large areas, especially in terms of variance, which would be expected to change depending on soil type (Voltz and Webster, 1990). Naturally, there are some specific problems with geostatistical methods. For instance they tend to smooth the appearance of classical soil maps and can fail to account for local variation. Conversely, however, the detail included in classical soil maps cannot be validated using the sparse sampling technique typically employed during the soil survey, hence polygon-based maps are produced using the soil surveyors' ill-defined tacit knowledge (Walter et al., 2006). A

lack of data is problematic for both approaches, as geostatistics require datasets to have a high sampling density (Lemercier et al., 2012), although a lack of empirical data will hamper every modelling approach to some degree. Both classification and geostatistical approaches have difficulty modelling sharp and gradual change in soil properties occurring in the same area.

1.2.3 Combining Approaches

As discussed, there are advantages and disadvantages to representing the soil as either a gridded surface or a set of polygons. Generally, a gridded representation can be used to depict local variation whereas a polygon will contain a lot more (tacit) information. As they are often required as an input for environmental or climatic models, the demand for spatially explicit soils data centres predominantly on soil property rather than class and hence a gridded output is generally desired. This is linked to technological advances in data capture and a rise in prevalence of mapping tools based in Geographic Information Systems (GIS) based which has prompted an increasingly sophisticated use of soils data (McBratney et al., 2003). Despite this, much of the available data on soils have been generated using traditional soil survey methods. This means that for many areas, the only soils data available are in polygon form. Moreover, point samples generated by the soil survey tend to follow a prescribed pattern. Where there are distinct chances in the soil forming factors, samples are generally located close to these 'boundaries' to confirm the surveyor's belief that the changes in soil forming factors have lead to a change in the soil class. In areas where there is perceived to be little change, there is generally a low sample density (Webster et al., 1979). This sampling scheme does not necessarily lend itself to the application of geostatistics (McBratney et al., 1981). Similarly, such an approach is usually not sufficient to ascribe soil property values to

polygons (soil classes) for use in other models, primarily due to uncertainty over the variability within and between polygons; especially as some properties will be weakly correlated with taxonomic classes (Heuvelink & Webster, 2001). Due to the constraints of data availability, combined with the limitations of both approaches, the most accurate predictions of soil properties are often made using a combination of gridded and polygon data, within a geostatistical model (Utset et al., 2000; Liu et al., 2006).

1.3 Digital Soil Mapping (DSM)

Understanding how soils can be represented is important as there is a growing demand for high resolution soil spatial data (Behrens & Scholten, 2006a). Globally, this demand stems from a desire to model phenomena such as, the propensity for flood generation and land management (Behrens et al., 2005). In the UK, the desire for soils data is motivated by issues such as the need to better understand the links between land management, runoff and water quality. For example, detailed soil information is required for the EU Water Framework Directive (Kallis & Butler, 2001), Catchment Flood Management Plans (Evans et al., 2002), the soil action plan for England (Defra UK, 2004) and the need to limit overgrazing in upland areas, plus UK agricultural policy (Mayr & Palmer, 2006). Although demand is increasing, collection of empirical data is waning due to the prohibitive cost involved and the time and manpower required. One solution to these conflicting problems is the use of Digital Soil Mapping (DSM) which, in the most basic sense, is a tool to create spatial information about soil (Behrens & Scholten, 2006b). DSM is a catch-all term for mapping soils using digital data. It encompasses many varied techniques which can be used to integrate point samples, gridded surfaces and class-based polygon maps to make predictions of soil and soil properties in whatever format is required (McBratney et al., 2003). Of the many DSM

methods available, this thesis focuses on employing those which make use of the soil-landscape mapping paradigm (Hudson et al., 1992). This is in order to maintain conceptual links to the original soil survey map production process, with a view to quantifying expert knowledge and integrating it into a DSM framework.

As data about the landscape have become easier to collect, store and manipulate through GIS and remote sensing, there has been a conscious attempt to quantify explicitly all the information, assumptions and approaches used to produce soil maps. Like the soil-landscape relationships used by the soil surveyors to produce soil classification maps, the underlying concept is that if the relationship between a soil and its environment is known for an area, it is possible to infer which soil type will be present at unsampled locations from associated environmental conditions (Zhu et al., 2001). A major difference between traditional soil survey mapping and DSM techniques is the treatment of uncertainty in the final soil map. One of the key theoretical underpinnings of DSM is the need to quantify uncertainty associated with predictions in a highly complex system such as soil (Heuvelink & Webster, 2001). This is especially relevant when information about soil is required in the context of wider environmental issues; for instance, when the uncertainty associated with the spatial prediction of soil organic carbon stocks needs to be quantified, as will be required if it is to be propagated into other models (Zhao & Shi, 2010).

The DSM approach provides a consistent and reproducible methodology which offers a clear measure of the error associated with each prediction. The initial stages of DSM involve the numerical or statistical modelling of the relationship between explanatory environmental factors and soil properties. When applied to a geographic database these

relationships can be used to produce a predictive soil map. According to Scull et al (2003) there are three main goals of DSM:

1. To use relationships between environmental and soil properties to collect soil data more effectively.
2. To produce a better representation of soil as a continuous, landscape variable.
3. To incorporate expert knowledge into predictive modelling.

DSM can be a purely data-driven exercise, using geostatistical interpolation between points. However, the integration of gridded geostatistical methods with soil classification data has been shown to improve the prediction of soil properties (Liu et al., 2006). In this thesis a predominantly knowledge-based approach has been adopted where, conceptually, the aim is to recreate a soil surveyors' thought process based on the principals of soil genesis (Rossiter, 2005), although purely empirical models have also been employed.

1.3.1 Data Sources

The reason direct observations of soil properties are scarce is due to the time, effort and expense required for their collection (Mayr et al., 2010). Despite this, many countries (such as the UK and Germany) have built up significant datasets detailing spatially referenced soils information – largely due to the efforts of their respective soil surveys. These datasets, in combination with advances in computing and statistics have allowed DSM to become the prevalent method of describing the spatial variability of soil and soil properties (McBratney et al., 2003).

1.3.1.1 Soil Legacy Data

In the soil survey, boundaries between soil classes are derived from changes in bedrock, Land cover and topography which are combined using the surveyors' training, knowledge and experience (Avery, 1980). This thought process relates back to ideas about soil genesis, or how soils are formed in the landscape (Jenny, 1941). Utilising these soil forming factors can provide a framework for predicting soil property and class (Lemercier et al., 2012). The environmental variables considered in the soil-landscape approach have been amended for DSM to include other soil properties (McBratney et al., 2003). Pre-existing soils data, typically in the form of soil maps (polygons) and the samples used to create these maps, is collectively known as legacy data. With limited funding for new data collection, legacy data has become an important resource in DSM (Mayr et al., 2010). Despite the aforementioned issues concerning the accuracy of polygon-based soil maps, legacy data can be a useful resource in the prediction of soil properties (Mayr et al., 2008), especially if used in combination with expert knowledge (which will be discussed in detail in the 'expert systems' section of this thesis)

The most detailed soil classification maps depict soil-landscape units most accurately and are, therefore, usually considered most useful for deriving rules linking soil properties to soil forming factors (Lemercier et al., 2012).. Generally, the desired scale of soil maps is 1:50000. However, this is normally available for only a small percentage of a countries land mass, if at all (Behrens et al., 2005). If the aim of the DSM exercise is to produce a soil classification map (which remains the most common method of displaying soil information) then it is difficult to make predictions from a limited number of quantitative samples alone (Mayr et al., 2010) because they often fail to represent the complete range of classes in the landscape. By using a pre-existing soil

map as the data source from which to develop predictive models of soil class, it is possible to create a large number of virtual samples using GIS, which can lead to more stable statistical modelling.

1.3.1.2 Ancillary Data

To conform to the soil-landscape approach, DSM requires data that represent climate, organisms (typically represented by Land cover or vegetation cover), relief (topography), parent material and time. There are many hundreds of metrics which can be used to represent these soil forming factors, which can be both grid- and polygon-based, and there is no definitive list of what should be used for different purposes (Bui et al., 2006). In many cases, the selection of variables used for DSM will come down to availability for the study area of interest. Incorporating ancillary data to the soil mapping process is desirable because the data are generally more readily available than quantitative soils data, either because they have already been collected (as with legacy data) or because they are cheaper to capture. Data describing topography and Land cover in particular will tend to be more prevalent than soils data due in part to advances in remote sensing technology (see 'Remote Sensing' section below).

The appeal of linking terrain attributes to the distribution of soils is that it is comparatively cheap and straight-forward to derive an accurate digital elevation model (DEM) using remotely sensed data. A large number of terrain attributes can then be derived from the DTM. Behrens et al. (2005) found that using a data mining technique (Artificial Neural Networks) in combination with a pre-existing soil map and 69 different geomorphic terrain attributes, it was possible to get reasonably accurate predictions of the spatial distribution of soil classes in the landscape. It should be noted that this prediction was improved by the inclusion of additional soil forming variables.

The importance of topographic predictors should not be overlooked, as the shape of the landscape plays a key role in a surveyor's decisions regarding soil class boundaries. Although there are a large number of terrain attributes which can be derived, is has been suggested that relatively few are required for the spatial prediction of soils. However, this will depend on the particular landscape under investigation (Mayr & Palmer, 2006).

As one of the major drivers for soil formation is the weathering of the rock *in situ* (Jenny, 1941), it is unsurprising that geology and parent material are frequently used as predictors. One issue with this is, much like the legacy soil maps, pre-existing maps of parent material can be quite inaccurate as they are typically mapped at low-resolutions (Mayr & Palmer, 2006). DSM studies using geology as a predictor have found that 'recent' Quaternary deposits (e.g. from the last ice age) are underrepresented and form the majority of 'missing' geological data (Lawley & Smith, 2008). This is an issue for the UK, because much of the landscape was glaciated and it is generally accepted that non-glaciated soils are more clearly linked to topography and surface geology than landscapes that have been produced by glacial deposits (Mayr & Palmer, 2006). Regarding the predictors used, the most relevant soil forming factors are likely to be scale-dependent. At a continental scale, climatic variables are likely to be more influential as opposed to much smaller scales where climatic differences will be imperceptible and local systematic variations will be mainly dependent on terrain (Bui et al., 2006). At scales in between, Land cover will often be a significant predictor, as the influence of human intervention, for instance, via ploughing, agriculture and irrigation will have a homogenising effect on soil properties (Webster, 2000). For this reason, studies into D_b are sometimes stratified by Land cover (Steller et al., 2008; Moreira et al., 2009).

1.3.1.3 Remote Sensing

The overview of the ancillary data available for DSM has highlighted the role of remote sensing for the rapid generation of comparatively low cost data, particularly for mapping the spatial distribution of land cover (Bossard et al., 2000). Moreover, advances in remote sensing technologies have enhanced both the amount of data which can be recorded and the actual data itself. Earth observation via satellite can provide continuous layers of information on spectral reflectance that can be related to both soil physical properties (e.g. particle size, surface roughness) and chemistry (e.g. organic matter content, mineralogy). Remotely sensed data is usually used in conjunction with ancillary thematic maps to improve predictive power. Hyperspectral mapping using numerous spectral bands can more readily identify individual minerals present in the soil, while radar and LiDAR are used to improve DTMs and habitat maps (Ben-Dor, 2002). As well as providing data used in DSM, remote sensing techniques have been applied to the measurement of soil properties directly.

Typically, airborne gamma radiometric data is used to detect changes in soil material; while this has proven to be successful in detecting broad spatial trends, its application for prediction is improved when it is used in conjunction with other ancillary data (Cook et al., 1996). This approach is especially of interest in remote areas, such as some parts of the Tropics, which are inaccessible and are often data poor (Minasny & Hartemink, 2011). The large volume, spatial coverage and low cost (in comparison to field sampling), make data generated using remote sensing of interest for DSM applications.

1.3.2 Soil-Landscape Statistical Models

The term ‘soil-landscape model’ has so far referred to a concept or paradigm regarding how soil variation is mapped. However, in terms of producing maps of the variation in soil properties across the landscape, the soil-landscape model refers to the formation of a statistical relationship between the soil and the soil forming factors. Soil-landscape models essentially try to replicate the soil surveyor’s mental or conceptual model. Rather than experience and training, data mining techniques are used to ‘learn’ rules dictating the spatial variation in soil- often using sophisticated algorithms (e.g. Behrens & Scholten, 2006b). The basic concept is that these data mining techniques or statistical models will develop relationships between environmental properties and a property of interest at a sampled location, which can be extrapolated to unsampled locations (Bui et al., 2006). Since including a range of soil forming factors within a single model requires a model to represent a multitude of undefined interactions between variables, it is therefore advisable to use an ‘adaptive, non-parametric model’. Two advantages of these techniques are that they adaptable enough to be applied to a range of problem solving applications and that they are able to model complex non-linear problems.

There are, of course, also difficulties associated with the soil forming approach to modelling, especially when it is applied to mapping soil classes. This is primarily because when pre-existing soil maps are used as training data, the probability of the class representing “real life” as it is mapped is never 100 percent. Maps have been smoothed out with the outliers removed and when testing the accuracy of predictions with measured data points, or extrapolating into unmapped areas, these outliers are reintroduced (Lemercier et al., 2012). The root of this problem is the variability within individually mapped soil units. This, combined with a low sampling density, a lack of a

definitive list of predictor variables representing the soil forming factors and interactions between these said variables makes prediction of class or attribute very difficult (Walter et al., 2006). These problems are just as applicable to the direct prediction of soil properties and even with modern statistical techniques that can handle the non-linear interactions between variables, there will inevitably be a significant amount of uncertainty that will need to be quantified.

Despite the difficulties, the key benefit of modelling using the soil forming factors approach is that the model starts to explain the variation in the soil, rather than just predict it. Even though the rules developed for mapping are extracted numerically from a dataset, the process can be seen as an important step in the process of knowledge discovery (Bui et al., 2006). Incorporating knowledge of processes into statistical models is intuitively desirable, but the parameterization is difficult (Heuvelink & Webster, 2001). This means that data miming will continue to remain an essential method for the future of digital soil mapping (Behrens & Scholten, 2006a).

1.4 Modelling Soil Bulk Density (D_b)

Statistical soil-landscape models are a fairly modern development in DSM because they were reliant on advances in computing capabilities (Heuvelink & Webster, 2001) and as such they have only recently been applied to the prediction of D_b (Martin et al., 2009). To illustrate the advantages of this approach for creating a landscape scale representation of the spatial variation in D_b, it is necessary to explain how D_b is typically predicted and how this affects the way it can be mapped.

1.4.1 Pedotransfer Functions (PTFs)

Pedotransfer Functions (PTFs) are defined as the "predictive functions of certain soil properties from other more easily, routinely, or cheaply measured properties" (McBratney et al., 2002) and their use builds upon the basic information collected during the soil survey, filling in the gaps between recorded information and information required for various predictive models. Originally developed for the prediction of soil hydraulic characteristics, their application has been broadened to include the modelling of chemical and physical properties (Wösten et al., 2001). As D_b is not routinely measured during the soil survey, it is frequently predicted from soil textural properties and organic carbon content (Adams, 1973; Rawls, 1983). Developing a statistical relationship between measured soil properties and D_b fixes predictions to the point (soil profile) scale, as there is no way of inferring D_b values in-between sample points. This can translate into a spatial prediction of D_b in two ways; firstly, if soil samples are taken on a regular grid, D_b can be derived from these properties and it can be assumed that the D_b values do not vary between points. This produces a gridded representation of D_b where each cell has a single D_b value (Bellamy et al., 2005). To capture the spatial variation of D_b across the landscape, this approach requires either a very large number of samples or the representation of soils at a very low resolution. Alternatively, PTFs can be developed for existing soil survey data and then an average value can be attributed to a soil class (Batjes, 1996). As stated, the problem with assigning a single property value to a class is that it disregards the (often significant) within-class variability.

The development of PTFs for D_b generally requires some pre-existing D_b data or sampling of new data as it is not advisable to apply existing PTFs to soils in any regions

beyond which they were developed (De Vos et al., 2005). In other words, the interactions identified by the PTFs between soil properties may not hold true across divergent soil series or heterogeneous geomorphic regions. PTFs are better suited to the prediction of soil properties (at the point scale) across small, homogenous regions where the relationship between the predictor and predicted variables is well established. They are less well suited to predicting across larger more heterogeneous regions; a problem which can be ascribed to the statistical technique of regression which is typically used to create PTFs. It should be noted that it is not the regression method itself that limits the prediction of PTFs to the point scale, it is the use of measured soil data. As there is no representation of soil data beyond the point at which it is measured, a model which uses soil data as an input can only predict at the point scale irrespective of which statistical technique is used. Despite this assertion, the regression method will limit the ability to model soil properties across a large heterogeneous landscape.

1.4.2 Regression

One of the most widely used approaches for modelling soil properties is the application of regression models, which are used to predict a response variable from explanatory variables. One advantage of these methods is that they can offer a clear set of descriptive statistics that measure the performance of the model, in terms of predictive error and standard deviation. Regression-based modelling approaches typically require continuous input data (i.e. not soil classes) and that the data are normally distributed (i.e. Gaussian) or have been transformed to be normal. Further assumptions such as those regarding multicollinearity and homoscedasticity are detailed in Berry (1993). Multiple linear regression has been used to model continuous soil properties (such as soil depth and water holding capacity) based on topographic indices with limited

success (Zidat, 2005). In this instance, the predictive capabilities of the models were greatly improved by the stratification of data into individual watersheds. This suggests that while topography alone was not able to adequately explain the variation of a number of soil properties, the use of additional landscape properties may improve model performance. One of the issues with regression models is that they generally require stratification as opposed to directly incorporating categorical landscape variables.

Regression models have been widely adopted and have been applied to the prediction of D_b in the UK (Hallet et al., 1998). Using samples taken from the Soil Survey of England and Wales (Hodgson, 1976), individual PTFs were developed using multiple regression of soil textural properties and organic carbon content, for individual horizons subdivided by lithological groupings of soil substrate material. Stratification had mixed results, with some horizons being predicted significantly more accurately than others; generally the A horizon was predicted with a greater accuracy than deeper-lying soils. To a lesser extent, the stratification by parent material lithology had some bearing on the predictive accuracy of the regression models. Disparity in the predictive accuracy of regression models is repeated in studies which stratify by soil classification (Calhoun et al., 2001). In these predictive PTFs, soil organic carbon (OC) content is consistently found to be the most important predictor variable.

For organic soil horizons, D_b is primarily determined by organic matter content as well as by how the soil was formed and the extent of humification. Linear regression analysis can be used to define the relationship between bulk density and organic matter content in these horizons (Hallett et al., 1998). Surface 'A' horizons are the most intensely weathered mineral horizons and hence the disturbance due to human activity is

considered as a predictor of D_b as well as organic matter content and particle distribution. Land cover is often employed as a proxy for human activity. Usually, the importance of organic matter content decreases with depth and particle size distribution becomes a more prominent predictor (Calhoun et al., 2001). Hallett et al. (1998) make the distinction in the 'B' horizon between podzolic or spodic soils from other mineral subsoils. For these horizons, parent material was considered to have the dominant influence on bulk density, hence, these soils were stratified by parent material group. Hallett et al. (1998) found that the pedogenic separation of data generally improved the predictions, in comparison to the running a regression using an unstratified dataset, with the improved performance attributed to the inclusion of *a priori* expert knowledge. Without applying some expert knowledge to a multiple linear regression model, it becomes very difficult to determine meaningful subsets of data to analyse. Generally, the inclusion of expert knowledge relates to choosing the correct soil forming factor(s) by which to stratify the dataset.

There are some problems with this approach, some of which will be regression specific, others will be more generic. A problem affecting all models concerns the input data. Measuring bulk density in the topsoil will only capture a 'snapshot' in the annual loosening and consolidation which occurs in cultivated soils; this might be particularly problematic in a country with a large fraction of agricultural land. Steller et al. (2008) highlight this issue in a regional study modelling D_b using soil chemical concentrations. They found that there is a weak relationship between topsoil organic carbon content and D_b across managed land, where the higher levels of predictive error can be associated with the regular occurrence of soil disturbance. Often, on agricultural land, cultivation and compaction by machinery will be the dominant influence, as opposed to soil OC

content and textural properties. This emphasises the need to account for various management techniques, especially over smaller spatial scales. Although, generally, stratification of data improved model performance, for some groups the regression models are very poor, offering little relationship between predictor variables and D_b. While in some cases this can be attributed to a lack of observed measurements for a particular soil taxon (Hallett et al., 1998), it highlights the variability of controls on bulk density within different soils across a single landscape, suggesting that for some soil horizons, organic matter, parent material and soil particle distribution may not be the dominant controls on D_b.

One potential problem with pedogenic stratification is touched upon by Heuscher et al. (2005) who suggest that certain soils will be favoured for agricultural use meaning they will tend to be intensely managed and, hence, will have a very mixed series of bulk density values as a consequence of cultivation. Predicting D_b using only OC and texture in these soils will tend to be difficult, therefore. Calhoun et al. (2001) believe that, even with stratification, many predictive models neglect the principles of soil genesis and morphology, only paying a cursory glance towards the influence of fundamental soil forming factors. Although D_b can be a function of depth, it is also clear that this relationship will change depending on parent material, vegetation cover, topography and internal drainage. To highlight the unpredictability of the variables which influence D_b, Benites et al (2007) found that total nitrogen was a significant predictor of Db in the Brazilian Amazon, whilst soil organic carbon content was not. However, when the data were corrected for co-variation, carbon was replaced by nitrogen as the most powerful predictor because soil carbon and nitrogen contents are strongly related. Benites et al (2007) explain that this could be because the Kjeldahl method for determining soil

nitrogen concentration is generally more precise than the method used for determining the concentration of carbon. This suggests that exploring the predictive power of other landscape factors not usually considered, is often worth investigation. However, it is possible that these lesser-used predictors may have more complex statistical relationships with D_b. In such cases, regression-based approaches would be less effective.

Pedogenic stratification of data sets does not always help to build more robust models, as Heuscher et al. (2005) discovered. Of the 48 sub-orders of soil examined , D_b was relatively poorly predicted in 13 ($R^2 < 0.40$) and reasonably well predicted in 14 ($R^2 > 0.60$) (note that the P-values for all models were <0.001). This relates back to the issues associated with soil class mapping, namely, that some classes are weakly correlated to certain soil properties (Heuvelink & Webster, 2001). The variation in the accuracy of prediction between sub-series is primarily based on organic carbon content, where soils with higher organic carbon are predicted more accurately. The study by Heuscher et al. (2005) found that, in the main, the D_b of younger soils and those occurring in a range of diverse locations (in relation to the suite of soil-forming factors) were the hardest to predict. Stratification by soil horizon has proved to significantly increase the power of predictive models, especially when the soils are further stratified by parent material. Despite this, there is a limit to how much stratification a single dataset can be subjected to and hence error associated with predictions may be due to unaccounted predictors or interaction between the existing model variables (Calhoun et al., 2001). As an aside, regarding legacy data, the Calhoun et al. (2001) study used data collected over a 45 year period, proving that historic data collected by many different soil surveyors can still provide strong results.

A major criticism of models used to predict soil properties is that they are often not tested on independent data sets and rarely tested using data from outside the specific ecosystem in which they were developed. This is particularly true of data intensive regression models. De Vos et al. (2005) addressed this by testing 12 published PTFs on an independent data set comprising 1614 samples of forest soils from Flanders, Belgium split into a two-thirds calibration:validation data set. Each PTF was applied to topsoil and subsoil separately. This study incorporates soil texture and loss on ignition (a proxy for organic matter) which were employed as predictors in accordance with the Rawls (1983) method. Again a measure of organic material was the most significant predictor of D_b. All models systematically failed to capture the full range of values, underestimating high D_b and overestimating low values. The study proposes that the D_b variation within forest soils is more pronounced as they lack the homogenising influence of agriculture.

1.4.3 Geostatistics

There are very few studies which have attempted to map the spatial variation in D_b using geostatistics. This is due to the fact that there is usually a lack of D_b measurements and that PTF derived point estimates of D_b are not suitable for interpolation. Kriging has been used to predict D_b successfully over very small scales for which a high sampling density is available (e.g. Entz & Chang, 1991). At a larger scale, Utset et al. (2000) found that using kriging was more accurate and less biased than predicting D_b using single values of soil properties which were associated with soil-class polygons. They also found that a combined soil map-kriging approach provided the most accurate predictions overall. It should be noted that the soil in this study area was relatively homogenous, containing only two distinct soil classes.

1.4.4 Measurement using remote sensing

Another approach to producing large scale maps of D_b is to measure the density directly using remote sensing technology, combined with a limited number of direct samples for calibration. However, Moreira et al. (2009) and Minasny et al. (2008) suggest that the power of this technique is currently relatively limited. Moreira et al. (2009) used near infrared reflectance spectroscopy (NIRS) as a rapid, low cost alternative to field sampling of bulk density measurements in the Amazon Basin, Brazil. Spectral reflectance was used to assess particle size distribution. However, in comparison with several published regression-based PTFs, it was found that the general regression-based PTFs overestimated bulk density and NIRS had a slight negative bias. Although data generated in this manner was not a powerful predictor for D_b, remote sensing has been applied to a range of other soil properties with varying levels of success (e.g. Ben-Dor, 2002) and can potentially provide useful input data for data-mining models.

1.4.5 Data-Mining Methods

The reason why regression methods have been explored in such detail is that many of the shortcomings associated with it are addressed using data-mining methods. Before the advances of modern computing, it was believed that representing soil forming factors in a predictive model was so complex as to be almost impossible (Kline, 1973). Over the intervening years, however, advances in GIS and data-mining have made this approach a feasible prospect. In comparison to both regression and geostatistical approaches, data-mining models have been shown to be those with the greatest predictive power for certain soil properties (Zhao & Shi, 2010). In various studies, two types of adaptive model are frequently seen to be the most robust predictors; Artificial Neural Networks and tree-based approaches (Behrens & Scholten, 2006a; Park et al.,

2005). There are other data-mining algorithms available, these two techniques were used in this study and, thus, are considered in more detail below.

1.4.5.1 Artificial Neural Networks (ANN)

Some fundamental problems with regression based models arose when large datasets containing vast arrays of digital soil data were developed. In such data, the influence of each predictor and the interactions between predictors were often difficult to identify using regression. To account for this, more sophisticated statistical models were developed, including ANNs which can be used to model more complex relationships by varying how predictor variables are connected to one another and the relative influence of each. ANNs are essentially a regression-based approach, with each variable represented by a node which is linked to other nodes by "weights" assigned empirically to represent the interactions between variables. During model training, these weights are systematically adjusted to best represent the relationship between predictors and the variable of interest (Behrens et al., 2005). Tranter et al. (2007) used an ANN to model bulk density using particle size distribution and depth as predictor variables in Australian soils. They found that D_b reaches a maximum when small silt and clay particles fill the spaces between sand particles. While D_b was found to have a logarithmic relationship with depth, the fractions of silt and clay sized particles appeared to have little or no predictive power. Tranter et al. (2007) found that ANNs overestimate low bulk densities and to underestimate high ones.

Calhoun et al. (2001) highlight the fact that, so far, it has only been possible to predict between 50-60% of the variance in soil D_b using organic carbon content and particle size distribution. This suggests that other further factors also control bulk density. These could include hydrological conditions and previous soil management, which

these models fail to account for. This is another compelling argument for the inclusion of a greater selection of predictor variables, encapsulating a wider range of soil forming factors. While ANNs have proved to be very powerful predictors of soil class (Behrens et al., 2005) and are frequently applied to regression problems regarding soil hydrology (Pachepsky et al., 1996) and chemical properties (Holmberg, et al., 2006), this method is not without its drawbacks. Primarily, the problems with ANNs stem from interpretation of model results. An ANN is a essentially 'black-box' model. This means that it is nearly impossible to assess specific relationships developed between the predictor variables and the final prediction (in other words it does not develop a set of rules linking the soil property of interest to the soil forming factors which can be interpreted: Mayr & Palmer, 2006). This means its potential for improving understanding is limited.

1.4.5.2 Tree-Based Methods

Classification and Regression Trees (CART), is a data-mining method developed to deal with very large (and, if necessary, incomplete) datasets containing both continuous, categorical and ranked data (Breiman et al., 1984). The technique has typically been used for ecological habitat mapping (Franklin, 1998) but can be readily applied to soil properties. Bui (2004) claims that as the prevalence of GIS and remotely sensed data layers increases, the classification of soil becomes more complex and so a decision tree model becomes more appropriate. CART (also known as decision tree analysis) can be used to extract rules from existing soil maps using DEMs and remotely sensed data sets as predictors. The CART model aims to partition data sets recursively into increasingly homogenous groups to determine whether groups of data differ significantly and which variables best explain these differences. Once partitioning is complete the subsets are

known as terminal nodes. The process is iterative from the complete data set (the root node), splits are dictated by how well they increase the purity of the dataset (homogeneity of the data being predicted). This splitting process outlines a set of rules that can be applied to other data sets. Usually, the tree will need simplification (called pruning) to avoid "over-fitting" to one data set (meaning the model describes noise in the dataset as well as the underlying relationships between variables). This is done by the cross-validation and amalgamation of terminal nodes. There are many different types of tree-based models, but all have the objective of producing a set of predictive rules developed from a training dataset that can be applied to other datasets (Scull et al., 2005).

There are several appealing aspects of CART, for DSM applications. One benefit is that, as a non-parametric method, it does not rely on a specific distribution of data, meaning data transformation is never an issue. In terms of modelling soil properties, linear models, such as the regression approaches discussed earlier, can fall short when there are several disparate landscape properties that support the same soil type. Bearing in mind that D_b values tend to fall into a relatively narrow range it is highly likely that, across a large, heterogeneous landscape, there will be a number of contrasting landscape characteristics that will contribute to similar D_b values. CART, on the other hand, can disentangle complex relationships between variables including interactions between predictor variables without *a priori* knowledge of the relationship (Martin et al., 2009). This is particularly useful when there are existing soil maps with no readily available soil point data, because samples can be extracted from digitised soil maps, helping to fill in gaps in existing soil surveys. Furthermore, these tree-based approaches do not suffer from the inclusion of data containing outliers or the inclusion of extraneous

variables (Friedman & Meulman, 2003). The principal drawback to this method is that there is no often clear single optimal model; adding and removing single variables from the model often affect model performance very little. This can make drawing conclusions from the model difficult since, by removing a single predictor variable, you can drastically alter the rules the model generates without impacting on the overall predictive accuracy. To counter this problem, Scull et al. (2003) suggest that, as a minimum, expert knowledge should be used to confirm that the splits made by the CART are sensible and that the optimal model derived has some plausible physical basis. CART has also been criticised for producing predictive maps with a stepped appearance; especially when the predictor variables are available at different spatial resolutions.

Recently, multiple additive regression trees (MART) have been used to predict D_b on a national scale. Martin et al. (2009) used MARTs to model D_b across France on the basis of organic carbon content, texture, land cover and depth. They found that the MART model was superior to traditional PTFs which were generally more biased towards the data mean leading to an overestimation of low bulk densities and an underestimation of high values. This suggests that PTFs may perform more poorly on large heterogeneous data sets compared to smaller homogenous sets, although Martin et al. (2009) purport that PTFs can be applied effectively across large scale data sets, so long as the data is partitioned into subsets. In the MART model organic carbon content was the most significant predictor of D_b and the qualitative variables had the lowest importance.

Martin et al. (2009) also propose a second MART model which had comparable results with the original model using just three predictor variables: organic carbon, clay and silt. The relative poor performance of soil type as a predictor was unexpected since soil

classification stores a lot of soil information. This suggests that, for D_b at least, there may be too much within-class variation to make soil maps a particularly powerful predictor. The main advantage of the MART model over traditional PTFs is the fact that it can better model the complex, non-linear interactions between variables (Martin et al., 2009). In other words, the MART model can separate the relationships between predictor variables (be they linear or non-linear) without *a priori* knowledge. By fitting numerous tree models, the MART method avoids the problems of limited predictive power associated with single tree models.

Two further methods which have been shown to improve the predictive power of tree-based models are Boosting (Elith et al., 2008) and Random Forests (Grimm et al., 2008). Essentially, both methods take the average predictions of many trees and are hence more robust predictors. The issue is that these models do not produce such a clearly defined set of rules as CARTs. In this instance, there is a trade-off between predictive power and interpretability. Tree-based models using multiple trees become 'black-box' modelling techniques (as are ANNs). This means it is nearly impossible to ascertain how the variables in the models are judged to be interacting and whether this conforms with present understanding of the soil-landscape concept. For this reason, the selection of which tree-based model to use depends upon the intended outcome of the study. One of the key motivations for using a CART modelling approach is the potential for knowledge discovery, which can be viewed as a method of uncovering and formalising the tacit rules contained in a map linking soil forming factors to soil class or attribute. These rules can then (in theory) be used to predict soil classes in unsampled locations (Hollingsworth et al., 2006). The CART method has been used to derive mapping rules based on geological and terrain attributes (Bui et al., 1999) and used to

predict continuous soil property data based on rules generated from pre-existing soul survey data (Henderson et al., 2005). The use of data-mining approaches to discover relationships between soil and various landscape attributes is clouded by model interpretability. To structure existing knowledge and provide a framework for knowledge-discovery, an alternative to data-mining, is an expert systems modelling approach.

1.4.6 Expert Systems

The numerical rules developed using data-mining, relate the soil to a suite of landscape variables. One of the key questions concerning this approach is whether the rules devised by these models are consistent with the principles of soil genesis (Bui et al., 2006). This is not straightforward, however, because the exact relationships between variables are often unknown due to the complex process involved. This raises the issue of including expert knowledge in predictive modelling. When a pre-existing soil classification map is used to develop the rules that are applied to unmapped areas, expert knowledge has entered the model implicitly. Rule extraction using CART formalises the decision process, and these rules can be judged by an expert to ascertain if they make pedogenic sense. Another approach is to generate these mapping rules using expert knowledge. Here, rather than simply interpreting the rules generated by a data-mining tool, the experts themselves dictate how changes in the landscape will affect the spatial distribution of soil classes or properties. This issue is explored by McBratney et al. (2002) who propose a shift from PTFs to soil inference systems. This approach to DSM attempts to integrate soil survey and mapping conventions and standards, incorporating soil surveyor's knowledge, while simultaneously reducing both cost and inconsistencies.

Inference engines differ from other models in two ways; they can utilise qualitative information not normally available in other statistical models and they have a unique structure known as a 'meta-model' which separates knowledge from the model. The knowledge generally takes the form of a set of rules which the model uses to make predictions. As these rules are clearly recorded, the process by which the model makes predictions is easy to interpret. The basic concept is that a soil map is "a structured representation of knowledge about soils' spatial distribution", hence they can be used *a posteriori* to establish rules regarding soil distribution (Bui, 2004). Knowledge is gathered from experts and is represented by a set of rules, procedures and logic. In this case, the soil surveyor develops general rules governing the relationship between the occurrence of soils and the landscape. This is conceptually quite similar to other modelling techniques as once again the soil forming factors and soil-landscape paradigms are used to organise knowledge of soil spatial distribution. The principal difference is that the knowledge is usually tacit, developed though the experience of the soil surveyor. In order to develop a formal model tacit human knowledge must be translated to a database which can be used by an inference engine and, from which, areas of similarity may be defined. This is known as knowledge programming. After a detailed survey, the rules are amended if required and rules for prediction are derived. Producing the final map is an iterative process, involving identifying exceptions to the general rules. Although expert knowledge-based systems are generally less data hungry than other modelling approaches, it should be noted that the development of such systems is much more challenging for areas where data availability is limited.

While the use of expert systems has clear potential for the predictive mapping of D_b, there are a few caveats to consider. First, the rules derived, and hence the model

produced, will only be as good as the knowledge available. This if the expert knowledge contains uncertainties, for instance in the landform characteristics that pertain to certain soil features, then the results will inevitably be inconsistent. A further issue relating to this is that common landscape descriptors used in DSMs, such as wetness indices and slope curvature, are not commonly used in manual soil mapping. They may, therefore, not be correctly attributed values when creating the knowledge base. There is also an issue around how the computer and the soil scientist handle data as the computer will always be consistent, hence reproducible. Conversely, tacit understanding may allow the soil scientist to identify some 'exceptions to the rule' based on local knowledge which can create a more accurate map. One of the major challenges that faces mapping using expert systems is how the technique is applied across larger, more heterogeneous landscapes, where the knowledgebase must be to be constructed by numerous soil scientists working in collaboration.

1.4.6.1 Bayesian Inference

Expert knowledge is built on a number of generalisations and is, therefore, subject to uncertainty. It is, thus, necessary to express the knowledge in a manner which can reflect this uncertainty. For this reason, expert systems sometimes utilise Bayesian inference, in which the relationships between variables are linked together in terms of probability (Jensen, 1996). While pure expert systems do not allow for the generation of statistical relationships between soil and landform to be generated through sampling, Bayesian modelling explicitly includes expert knowledge within the statistical relationship. This method is flexible because it can be used as a data-mining tool to predict soil class (Mayr et al, 2008; Mayr & Palmer, 2006), or it can incorporate expert-

derived mapping rules to predict soil class (Skidmore et al., 1996; Bui et al., 1999) or soil attribute (Corner et al., 2002).

Bayesian modelling can provide expert knowledge with a quantitative framework, which allows errors associated with predictions to be expressed numerically. This is important if it is going to provide a feasible alternative to data mining, geostatistics, PTFs and traditional survey. The use of expert knowledge may be particularly applicable to the prediction of soil properties, as there are typically less data available and hence a greater potential to 'fill in the gaps' regarding prediction. As stated earlier, within any mapped soil class, there is a degree of variability regarding the soil properties within. To represent this, Corner et al. (2002) developed a set of predictive rules to map the surface clay content of Australian soils which, in terms of accuracy, improved upon a traditional map-based estimate. Moreover, by separating the clay content into classes, it was possible to produce an associated probability map stating the likelihood that a mapped attribute fell within a predefined range. Since a limited amount of data was available for this particular study, the probabilistic relationships between variables were derived from coincidence matrices sampled using a Bayesian GIS tool called 'Expector' (Cook et al., 1996). These initial estimates of joint probabilities were then amended by an expert. As there were limited empirical data available, the expert adjusted probabilities were considered less likely to be bias in comparison with those derived from data alone.

There are a number of drawbacks to this approach, one of which being that continuous datasets must be discretized (changed from a continuous range to a set of distinct classes). For example, if slope gradient is a predictor, rather than being input into the model as a continuous variable e.g. 0-30°, it needs to be classified e.g. 0-10°, 10-20°,

20-30°. This is potentially problematic as it is hard to define meaningful boundaries between classes while simultaneously keeping the total number of classes low enough to make the model usable for an expert. While this is not the case for all Bayesian methods (Lunn et al., 2000), it remains a common problem. Although there are no definitive guidelines regarding the classification of landscape attributes into formal geomorphic units, it is a problem that can be circumvented using expert knowledge. By allowing the expert to define the boundaries, it is possible to create classes that are more meaningful and relevant to the soil surveyor (Corner et al., 2002). Continuing the slope gradient example, the expert may determine that for the soil property being predicted, the most meaningful class boundaries are 0-3°, 3-10°, 10-30°. Furthermore, most data-mining approaches also partition input data in some manner, so that discretization issues are not confined to Bayesian modelling.

Another problem, although again not one unique to Bayesian modelling, is that predictive accuracy is limited by the accuracy of input data. For example, the soil maps used as the input data are representations rather than accurate depictions of reality. This means that error will be introduced into spatial predictions due to the fact that the map is incorrect rather than the model. It is possible to address this difficulty using Bayesian methods, by giving each class in the input data (in this instance the soil map) a probability that it is realistic. While this does not solve the problem, it does account for the uncertainty the problem creates.

A further concern associated with some Bayesian expert systems used to date (Cooke et al., 1996; Corner et al, 2002) is that the method relies on the assumption that input evidence layers are independent. In reality, due to the nature of spatial data, there is likely to be some dependence especially for variables used to predict the same property.

While this is not necessarily problematic, especially in the case of weak interactions or when the statistical correlation does not reflect the same pedogenic processes, it may need to be addressed. One solution is to use Bayesian networks (BNs) as a modelling tool, as due to their model structure, the independence of input data layers can be addressed using the assumption of conditional independence (Mayr et al., 2010). The theory and application of Bayesian networks will be discussed in more detail in Chapter 3.

1.5 The ISIS Project

European policy directives, in particular, the Thematic Strategy on Soil Protection (COM(2006)231) and the proposed Soil Framework Directive (COM(2006)231), are an attempt to legislate against soil degradation and promote better management (INSPIRE EU Directive). One of the data requirements needed to support this legislation is a national-scale 1:250000 soil map. This led to the development of the Irish Soils Information System (ISIS) which will create this map and the soil information contained therein for the Republic of Ireland. Beyond creation of an Irish soil map, ISIS aims to support ongoing soils research by incorporating information regarding soil quality and soil functions. Moreover, this information needs to be available in a usable format and accessible to a range of interested parties, including scientists, policy makers and the general public (Daly & Fealy, 2007).

Before ISIS, pre-existing detailed soil information was available for just 44 percent of the country; which was mapped at a scale of 1:126,720 by An Foras Talúntais (Now Teagasc - The Agriculture and Food Development Authority) in a soil survey conducted between 1959-1985 (Gardiner & Ryan, 1969). The soil maps produced from this survey are detailed to soil series level and include associated profile descriptions which provide

typical values of a number of soil properties associated with the series. Beyond this, there is a generalised soil map at a 1:575,000 scale which shows soil associations at Great Group level (groups of soils related to particular landscape features) rather than individual soil series (Gardiner & Radford, 1980). This means that for the 56 percent of the country that has not been surveyed, there is a general account of the types of soils likely to be present, but without associated soil property information.

The basis of the ISIS soil classification is the soil series. These series are described as the 'information carriers' of the project (Daly & Fealy, 2007), meaning the classification differentiates soil attributes on the basis of soil series. The challenge for ISIS is to map the series into areas which have not been surveyed. To accomplish the task of completing a 1:250000 scale soil series map of Ireland, the project uses digital soil mapping techniques (McBratney et al., 2003) to combine data from the pre-existing soil maps with new data in the form of approximately 300 detailed soil profile descriptions and several thousand auger points across the country. While this mapping is in progress, and new data is being generated, it is the intent of this PhD to investigate some of the methods of populating the newly defined soil series with additional soil data; in this instance, regarding D_b.

1.6 Gaps in Knowledge

This literature review has identified a number of research opportunities surrounding DSM in general and the prediction of soil D_b specifically. McBratney et al. (2003) note that of the studies predicting either soil class or attribute, most are small scale. Of approximately 70 studies they examined, the median study area was 30 km², with pixels usually representing around 20 m. Scull et al. (2003) encourage future research at a larger scale and modelling at a national scale is clearly a different challenge because at

this scale, it can be argued that developing a single PTF from a data set collected from across the country is probably not a suitable method. Most existing models predicting D_b over large geographical regions tend to underestimate high values and overestimate low ones (De Vos et al., 2005). The fact that predictions tend towards a central mean should not be surprising as at a national scale, it is probable that the model will span several disparate regions. The use of a PTF in these circumstances will average the effects of a variety of different dominant processes controlling D_b, which does not make pedogenic sense.

A significant amount of work predicting soil properties has been conducted in arid areas, often with fairly topographically uniform landscapes, which can be considered more straightforward to model, as fewer factors will control variation in soil properties (Scull et al., 2003). Although, a potential advantage of modelling in more temperate climates is that that seasonal changes in D_b due to wetting and drying cycles reported in Pires et al. (2009) may be less pronounced. This is partly because even in the summer months there are rarely prolonged dry spells and partly, as Lee et al. (2009) explain, that variation in D_b caused by a wetting-drying cycle is most prominent in soils with vertic properties, i.e. ones where the subsoil contains 35% clay or more, of which there are relatively few. Arrouays et al. (2009) identify two significant areas warranting further investigation, which this thesis hopes to engage with. One is the representation of uncertainty within the modelling process and the other is the development of new techniques and data sources, as potential model inputs.

With regards to D_b, one of the knowledge gap is whether continuous landscape attributes can be used to model D_b data mining techniques. At present, the use of data-mining techniques for the prediction of D_b has mainly been restricted to using point

samples of soil properties, which limits prediction to the point scale (Tranter et al., 2007; Martin et al.; 2009). Zidat (2005) attempted to create a continuous map of soil properties using terrain attributes and regression modelling and found the predictive accuracy of the results to be very poor. The potential benefit of data-mining techniques is that they can incorporate more data (including categorical data) meaning more of the soil forming factors can be used for modelling. Moreover, they are capable of representing complex, non-linear interactions between a range of variables, suggesting that the predictive accuracy of data-mining models for soil properties may be an improvement on regression models.

Another question is how gridded predictions of D_b can be used as inputs of other models, in particular those used to predict soil carbon stocks (Grimm et al., 2008). This is of particular interest in comparison to how results compare with the use of an average D_b for a soil class, which is the current situation.

Finally, significant knowledge gaps exist in the use of expert systems and expert knowledge for soil mapping. Although Bayesian networks are established in ecological modelling (Kuhnert et al., 2010) they are rarely used for DSM. There is considerable potential for research into the ability to explicitly include expert knowledge within the modelling framework to improve statistical predictions. As a natural extension to this, there is scope to investigate the extent to which expert knowledge can be a stand-alone resource for DSM of soil properties.

1.7 Aims and Objectives

1.7.1 Aims

This thesis has two high level aims:

1. To investigate the utility of soil-landscape models to produce a spatially explicit map of D_b
2. To demonstrate that expert knowledge can be used to improve soil-landscape models for the prediction of D_b

1.7.2 Objectives

In order to achieve these aims, the following objectives need to be achieved:

1. To develop a set of PTFs for soil D_b using a range of data-mining techniques developed from soil textural properties and organic carbon content and then to attempt to improve predictive capabilities by including a range of soil-forming landscape attributes
2. To test whether soil-forming landscape-scale variables alone can be used to predict D_b and, from this model, to produce a spatially explicit map of D_b
3. To demonstrate the importance of a spatially explicit representation of D_b in relation to the development of soil carbon stock inventories.
4. To test whether a Bayesian Network can be used as a suitable data mining tool to predict D_b
5. To show that incorporating expert knowledge in the model structure of a Bayesian Network can improve the accuracies of prediction.
6. To develop a naive Bayesian network to predict D_b using expert knowledge as a proxy for data
7. To develop an expert structured Bayesian network to predict D_b using expert knowledge as a proxy for data

8. To populate a soil taxonomic system with D_b values generated using data-mining and expert knowledge-based predictions and to compare the results to the reference values used for soil series

1.8 Thesis Structure

In order achieve the objectives of the project, the thesis is divided into three experimental chapters. Chapter 2 uses data from a study area in the UK to address Objectives 1-3. Chapter 3 uses the same data and UK study area to address Objectives 4-5. Chapter 4 uses a different set of data, collected from a study area located in Ireland, to address Objectives 6-8. Each of these experimental chapters includes an independent methodology. In Chapter 5 the key findings of each experimental chapter are discussed, along with the implications of these findings and a perspective on future work which is required in this area. Chapter 5 also highlights how the thesis has made an original contribution to knowledge and offers a conclusion for the thesis as a whole.

1.9 Publications

In addition to the thesis the following publications have been produced:

- Taalab, K. P., Corstanje, R., Creamer, R. and Whelan, M. J. (2013), Modelling soil bulk density at the landscape scale and its contributions to C stock uncertainty, *Biogeosciences*, vol. 10, pp. 4691-4704.

- Taalab, K. P., Corstanje, R. Zawadzka, J., Mayr, T., Whelan, M. J. and Creamer, R. (2013) On the Application of Bayesian Networks in Digital Soil Mapping (*in review*)

- Taalab, K. P., Corstanje, R., Mayr, T., Whelan, M. J. and Creamer, R. (2013), The Application of Expert Knowledge to Bayesian Networks to predict Soil Bulk Density at the Landscape Scale (*in review*)

2 Modelling Soil Bulk Density at the Landscape Scale

This chapter tests the statistical modelling methods multiple linear regression (MLR), artificial neural networks (ANN) and Random Forest (RF) for the creation of PTFs to predict D_b. These PTFs use soil textural properties and OC content as predictor variables. The next set of models developed, include a suite of landscape variables, in an attempt to improve predictions. A third set of models, which do not rely on measured soil properties as predictors, are then tested. The aim of the third set is to use predictor variables which are not fixed at the point scale (soil texture and OC content), and hence create a gridded, landscape-scale prediction of D_b, which is required for a number of modelling applications. To demonstrate the advantage of using a gridded prediction as opposed to the soil class-mean D_b commonly used, the two approaches were used to calculate soil carbon stock across a study area in the Midlands, UK. Over the entire study area, the total stock inventories were similar, however, the error associated with the stratified mean D_b predictions were nearly twice as large as those of the gridded model. At a smaller scale, there was more variation in stock estimates between the methods, with a difference of nearly 15 t ha^{-1} of carbon in some regions.

2.1 Introduction

Bulk density (D_b) is defined as the oven-dry mass per unit volume of a soil (IUSS Working Group, 2006). It is a property of the soil which is typically predicted, rather than measured, as to do so is costly and time consuming. The most frequently used method of prediction is a pedotransfer function (PTF) which infers D_b values from other more routinely measured soil properties, usually soil textural properties and organic carbon content (Rawls, 1983). This modelling approach fixes D_b at the point scale, the scale at which the predictor variables are measured. While PTFs relate soil properties to

other soil properties, the influence of the soil-landscape approach to modelling can be accounted for in the stratification of datasets. The soils data used to predict D_b can be stratified on the basis of Land cover (Steller et al., 2008), parent material (Calhoun et al., 2001), soil class (Heuscher et al., 2005), horizon (Hallett et al., 1998) or a combination of several (Hollis et al., 2012). Stratification is necessary if landscape variables are to be accounted for when using a PTF, as the technique tends to be a form of regression model, hence not designed to include categorical variables. Recently there have been attempts to explicitly include landscape variables as predictors of D_b, using a tree-based modelling approach (Martin et al., 2009). Table 2-1 shows the results of a number of previous studies which have predicted D_b at a landscape scale. These studies still generate point scale predictions, as measured soil properties are used as predictor variables. From a point measurement or prediction, the spatial variation of D_b across a landscape can only be represented as an average within a fixed area, usually a soil class (Batjes, 1996) or a low-resolution grid (each cell representing several km²) based on a regularly distributed set of sample points (Bellamy et al., 2005). The soils data required to solve numerous environmental problems are generally high-resolution gridded data (Behrens & Scholten, 2006a).

Generating data about D_b is important for a number of reasons, for example, it dictates water and solute movement through the soil, can be indicative of soil quality for agriculture and is vital for soil carbon and nutrient stock assessment (Bellamy et al., 2005; Ungaro et al., 2010; Martin et al., 2011). The focus of this chapter is on how D_b can affect estimates of soil carbon stocks. This is a highly important area of research as after the oceans, terrestrial ecosystems are the second largest store of carbon on earth, with the majority contained in soils (Batjes, 1996). These terrestrial carbon pools are

highly susceptible to short term variation and are readily affected by anthropogenic influences such as Land cover change. Consequently, they play a critical role in determining current and future global carbon budgets (Bellamy et al., 2005). Soil can either be a net sink or source of carbon (Janssens et al., 2005) and there is continuing debate as to its potential to mitigate atmospheric CO_2 emissions (Smith et al., 2005). The accuracy of soil carbon stock estimations is, therefore, of paramount importance.

Dawson and Smith (2007) suggest that much of the error associated with carbon stock inventory in soils, can be traced back to uncertainties in D_b estimates, prompting further investigation into the methods for deriving these estimates. Specifically, errors are attributed to using mean values of D_b across large regions (often stratified by soil type) rather than a spatially explicit representation. Furthermore, soil carbon content plays a crucial role in spatially distributed, integrated land-atmosphere process models such as JULES (Harrison et al., 2008). There is evidence that improvements to the soil C component in these type of models increases their response-sensitivity to changes in soil stocks and processes. For instance, Jones et al. (2005) compared the outputs of a non-distributed soil C model to those from a multipool, distributed, soil C model and found that there was a difference in the magnitude of the feedback between climate and soil C when the distributed model was considered. Estimating the size of spatially distributed carbon pools requires a spatially distributed estimate of D_b.

Table 2-1: Results of previous landscape-scale bulk density predictions. *SRMSE- scaled root mean squared Error. **De Vos et al. (2005) tested 12 published PTFs on independent data from forest soils. Only the best performing model is reported

Statistical method	No. Samples calibration/ validation	RMSE	R^2	Predictors	Stratification	Location	Scale	Authors
Stepwise multiple regression	1396/ N/A	0.19	0.66	Clay content, TOC, sum of basic cations	Soil depth, soil order	Rio de Janeiro, Brazil	Regional	Benites et al. (2007)
Stepwise multiple regression	937/ N/A	Not reported	0.72	SOC, sand, silt, clay, physiography, parent material, horizon, vegetation, texture, consistence, structure, Munsell colour, drainage.	Parent material	Ohio, USA	Regional	Calhoun et al. (2001)
Non-linear regression	N/A/1614 (validation only)	0.45	0.59	Natural-log-transformed organic matter	Top soil/ sub soil	Forest soils, Flanders, Belgium	Regional	De Vos et al. (2005)**
Multiple regression	1568/ N/A	0.100*	0.77	OC, clay, silt, sand	Soil horizon, lithological groupings of substrate material, Land cover	England & Wales	National	Hallett et al.(1998)
Stepwise multiple regression	46987/N/A	0.188	0.45	Silt, clay, sand, water content, depth, organic carbon	Soil sub-orders	USA	National	Heuscher et al. (2005)
Multiple regression	1545/818	0.16	0.56	OC, clay, sand	Land cover, Soil class, Horizon	Europe	Continental	Hollis et al. (2012)
stepwise multiple linear	337 topsoil/ 1283 subsoil	0.10/ 0.14	0.72/ 0.62	Particle size classes (Swedish standard)	Top soil/ sub soil	Agricultural soils, Sweden	National	Katter et al. (2006)

regression								
Boosted regression trees	3131/x-validation (at a 0.9:1.0 ratio of training to test data)	0.0489	0.666	OC, silt, clay, gravel, soil, land cover, depth layer	N/A	France	National	Martin et al. (2009)
Cubist regression tree	93/39	0.09	0.26	mid-infrared diffuse reflectance spectroscopy	N/A	New South Wales, Australia	Regional	Minsay et al. (2008)
Multiple regression	1184/x-validation (model calibrated on <200 samples)	0.13	0.14	Spectral reflectance	Land cover	Amazon Basin, Brazil	Regional	Moreira et al. (2009)
Exponential regression	146/NA	0.158	0.896	SOC	Land cover	Southern Wisconsin, USA	Regional	Steller et al. (2008)
Multiple linear regression + expert knowledge	357/189	0.176	0.549	Organic matter, depth	N/A	Australia	Regional	Tranter et al. (2007)
Artificial Neural Networks	357/189	0.155	0.480	Sand, Organic carbon, depth	N/A	Australia	Regional	Tranter et al. (2007)

There are two principal approaches to estimating carbon stocks. One is to predict soil carbon concentrations across the landscape (often using geostatistics) and then combine these with a measure of D_b and depth to predict the stock (Ungaro et al., 2010). The problem with this is that using mean D_b values to convert predicted soil OC concentrations into OC stocks (i.e. the failure to use spatially varying D_b values) is flawed because important variations within individual soil types are omitted (Grimm et al., 2008). Alternatively, stock can be predicted directly across the landscape (Jones et al., 2005). The issue with this approach is that it cannot account for variations in the relationship between OC content and D_b across the landscape, fixing this relationship at the point scale. This makes prediction at the landscape scale difficult, as at that scale, soil properties are driven by physical environmental gradients and boundaries, such as topography, parent material and hydrologically effective rainfall. One of the most important recent research themes of international interest is the anticipated change in terrestrial carbon stock under changing climate and land-use (Yu et al., 2012, Zaehle et al., 2007). By modelling D_b using these changing landscape attributes, it can be viewed as spatially variable rather than as a fixed soil property. This may be an important consideration when predicting changes in soil carbon stocks over time, as both the soil carbon concentration and D_b will vary with changes in climate and Land cover. Lastly, large datasets containing measurements of soil properties are scarce, prompting investigation into the possibility of making predictions using landscape variables.

Soils are formed through the combined effect of physical, chemical, biological and anthropogenic processes on soil parent material. These factors will affect soil formation in different ways across the landscape, resulting in the spatial variation observed in D_b. Defined originally by Jenny (1941), these factors are; soil, climate, organisms, relief,

parent material, age and landscape position (SCORPAN). Today this information can be obtained from existing, large-scale soil maps, climatic data, Land cover maps, digital terrain models and their derivatives, parent material/geology and landscape position. The relationship between measured D_b and the soil forming factors at the sampling location and in the surrounding landscape can be formalized using statistical models (McBratney et al., 2003). These models are developed based on existing data and expert- or empirically-derived soil-environmental relationships. They can then be used to predict D_b within a landscape.

Recently, these principals have been applied to the prediction of both D_b (Jalabert et al., 2010, Martin et al., 2009) and organic carbon stock (Wiesmeier et al., 2011, Grimm et al., 2008) at the point scale with considerable success. Methods commonly used to explicitly include landscape attributes in the modelling process are Artificial Neural Networks (ANNs) (Keshavarzi et al., 2010) and Random Forests (Prasad et al., 2006).

The objective of this study is to determine the efficacy of soil landscape models to predict D_b for any given landscape, using a range of models. The intent is not to determine the best modelling method, but rather to cover both linear (MLR) and nonlinear (Random Forests and ANN) predictive methods to establish the feasibility of a landscape level prediction of D_b. This study considers both data rich (including measured soil properties) and data poor environments (models which do not include OC or soil textural properties as predictors) in which prediction is reliant on landscape-derived attributes. This allows the production of spatial estimates of D_b without interpolation and which prompts discussion about the implications of spatial uncertainty for the wider modelling community.

2.2 Materials and Methods

2.2.1 Data

2.2.1.1 Soil Survey Data

The soils data for this study were obtained from samples collected between 1970 and 1987 during the 1:25000 and 1:50000 soil mapping of England and Wales. The dataset has been described in detail by Hallett et al. (1998). Undisturbed 222 cm^3 soil cores were taken in triplicate using the methods detailed by Hodgson (1976), the D_b and other soil measurements (organic carbon content, particle size fraction, textural class and depth of the horizons) were derived using methods described by Avery and Bascomb (1982). Due to limitations of computational power required to derive landscape attributes for the whole country, a subset of the data was selected from a limited area (a 18150 km^2 region of the English Midlands) based on the relatively high density of samples (Figure 2-1). One issue which must be addressed is the locations of the sample sites, as they were not selected using a recognised sampling scheme. Samples were taken as part of the soil survey of England and Wales, meaning the locations of the sample sites were at the discretion of the soil surveyor. In areas where the soil was perceived to be relatively uniform by the surveyor, there will typically be few samples, whereas, in areas that marked the boundary between soil groups, there will often be numerous samples which were used to confirm the surveyor's hypothesis (Clarke, 1940). This inevitably leads to bias in the dataset, however, as resampling is not a feasible option, it is sufficient to be aware of this potential source of bias. The soils in the area are dominated by brown earths and surface water gleys, most of which have either a coarse or fine loamy texture, with some more clayey soils in the south of the region (McGrath & Loveland, 1992). The bedrock is dominated by undifferentiated

argillaceous rocks with prominent areas of sandstone in the west and patches of limestone in both the north and south. The elevation ranges from -2 m to over 550 m. The spatial representation of soils data comes from the National Soil Map of England and Wales (NATMAP: Hallett et al., 1996) which is a 1:250,000 scale soil map classification map. The classifications used in this study were at the Association (many, homogenous groups) and Great Group (few, more heterogeneous groups) levels. Here soils are differentiated on the basis of particle size and organic composition, calcium carbonate content and mineralogy (which affects nutrient supply and can be primarily attributed to the soil's parent material (Avery, 1980).

The environmental covariates which make up the predictor variables of each model were selected on the basis of three criteria: availability, cited literature and expert knowledge. Firstly, the data needed to be available for free as there was no budget to buy data. This means that often the data is of lower resolution that would ideally be desired (e.g. land cover). There are a huge number of environmental covariates which could be used, however, this study focused on those which have proved to be influential in previous studies (Behrens et al., 2006a; Grimm et al., 2008; Martin et al; 2009). Finally, expert opinion was used to select other variables (such as SAGA wetness index and sediment transport index) which could feasibly have an effect on or relationship with D_b.

2.2.1.2 Topographic Data

Although not usually applied to the modelling of D_b, topographic model parameters are frequently used in digital soil mapping (McBratney et al., 2003) and have been specifically used to predict soil organic carbon concentration (Grimm et al., 2008). A 10

m resolution digital elevation model (DEM) (Figure 2-1b) was used to derive the following landscape attributes: elevation, slope, aspect, curvature (plan, profile and mean), The DEM is accurate to ±2.5 m and was derived using photogrammetric methods by the Ordnance Survey, UK. SAGA wetness index (SWI) and sediment transport index (STI), all of which are commonly used topographic features in digital soil mapping (Wiesmeier et al., 2011). The SWI is based on the ratio of contributing upslope area per unit contour width and local slope angle (Böhner et al., 2002). The STI is based on unit stream-power theory, where upslope contributing area is directly related to discharge (Moore & Burch, 1986). Classification algorithms were used to divide the landscape into 7 and 8 homogeneous topographic classes on the basis of curvature, slope and catchment size (Pennock et al., 1987), and slope, surface texture and local convexity, respectively (Iwahashi & Pike, 2007). The derivation of these landscape attributes was carried out in ArcGIS 9.3 (ESRI, 2009).

2.2.1.3 Climatic Data

The following climatic data were used as predictor variables: average annual rainfall (mm yr^{-1}), accumulated temperature above 0°C, median number of field capacity days (i.e. the number of days per year that the soil moisture content is above field capacity), annual average potential evapotranspiration (mm yr^{-1}) and maximum potential soil moisture deficit (i.e. the water required to bring the whole soil profile back to field capacity, mm). The data were originally derived as 1971-2000 averages from monthly reports by the UK Meteorological Office, which provides information on weather for a 5 km x 5 km grid (Perry & Hollis, 2005). Average annual rainfall is the total of the monthly means per year and the accumulated temperature above 0°C gives an effective daily temperature above 0°C per month (Hallett & Jones, 1993). Evapotranspiration was

calculated using the Penman-Monteith equation, as detailed in Hess (2000), while the potential soil moisture deficit (based on the balance of rainfall and evapotranspiration) was calculated using methods described by Jones & Thomasson (1985). Field capacity days is the median number of days per year that each soil type is calculated to be at or above field capacity based on water balance calculations (assuming free drainage) over the period 1970-2000 (Jones and Thomassen, 1985).

2.2.1.4 Geology

Soil properties derive, in part, from *in-situ* weathering of the parent material (Grimm et al., 2008) so a representation of geology is essential for a digital soil mapping approach. A 1:50000 geological map was obtained from the British Geological Survey (BGS) which included the rock lexicon, giving the major rock unit (available for download from http://edina.ac.uk/digimap) and the BGS rock classification scheme detailing the lithology of the bedrock. The distribution of bedrock, by rock classification scheme, is shown in Figure 2-1c. The same classification scheme was also used to categorize superficial deposits (formerly known as drift), which represent the most recent geological deposits. Parent material was represented using the NATMAP 1:250,000 soil map (Hallett et al., 1996).

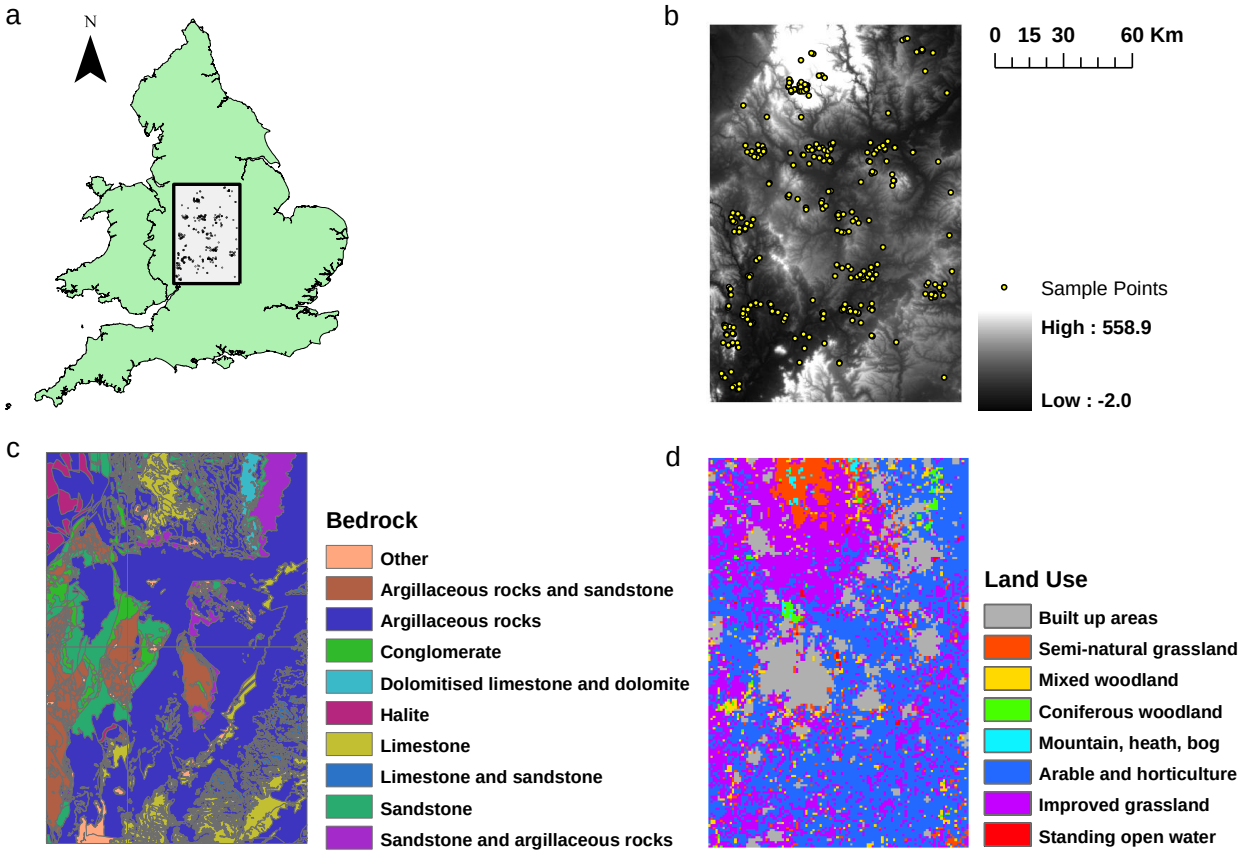


Figure 2-1: Location and study area. a) Study location in relation to England and Wales. b) Digital elevation models and sample locations. c) Geological rock classification scheme. d) Dominant Land cover classes.

2.2.1.5 Land Cover

The land cover (Figure 2-1d) was represented by the Land Cover Map 2000 produced by the Centre for Ecology and Hydrology (CEH). This map was re-coded to reflect the Land cover at the time of the bulk density sampling (which was recorded at the time of sampling). Satellite imagery was classified into a 25 m raster dataset which was subsequently aggregated to a ten-class 1 km grid land cover map (Fuller et al., 2002). Previous studies have commonly only attempted to make predictions within a single Land cover such as agricultural soils (Katterer et al., 2006) or forest soils (Jalabert et al., 2010). When the region is heterogeneous, Land cover has proved to be an important determinant of D_b (Hallett et al., 1998; Moreira et al., 2009). In this case, as Land cover was recorded when the D_b samples were taken, the land cover map was re-coded to reflect changes over time.

2.2.1.6 Soilscapes

To help evaluate the spatial performance of the models, results are assessed by "Soilscape". Soilscapes are landscape units derived from expert knowledge based on the 300 soil associations that make up the National Soil Map (Soil survey of England and Wales, 1983; Mackney et al., 1983). Each association has a subgroup code (Avery, 1980) that identifies the diagnostic soil properties. From this, the Soilscapes have been delineated by agglomerating National Soil Associations resulting in 25 classes. Within these national classes, the Soilscapes have been subdivided and grouped into homogenized regions based on similarities in drainage characteristics, texture and geology (Farewell et al., 2011). A description of predictor variables used in this study,

including their derivation and the number of classes or range of values in the study area is shown in

Table 2-2. It should be noted that stratification by Soilscape may not produce the best results and that a data-driven stratification based on different spatial layers could be explored in order to avoid the influence of subjectivity in the delineation of soilscapes. Despite this, assessment by Soilscape is deemed to be adequate for the purpose of this study as it will give a measure of model performance by expert-defined categorisation, which is one of the ultimate aims of the work.

Table 2-2: Predictor variables used in the ANN and RF model. The variables are listed in order of importance for the RF model predicting A horizon D_b.

Name	Description	Number of classes/ Range
Land cover	Land cover derived from the 1 km x 1 km Land Cover Map 2000 produced by the Centre for Ecology and Hydrology (CEH) (Fuller et al., 2002)	14
Soil Association	Soils grouped to the association level (Avery, 1973) derived from a 1:250,000 scale National Soil map of England and Wales (NATMAP; Hallett et al., 1996).	24
Elevation	Elevation above sea-level derived from a 10m DEM (Childs, 2004)	-2 - 558.9 m
Great group	1:250,000 scale National Soil map of England and Wales (NATMAP; Hallett et al., 1996) classified into soil Great Groups (Avery, 1980)	5
AT0_Annual	Average accumulated temperature above 0°C derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry &	2564 - 3871 °C

	Hollis, 2005)	
Parent Material	Soil parent material derived from a 1:250,000 scale Soil map of England and Wales (NATMAP; Hallett et al., 1996)	18
PSMD	Potential soil moisture deficit related to the balance between rainfall and potential evapotranspiration (Jones and Thomasson, 1985) derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005)	50 - 261 mm
PT	Potential evapotranspiration is the amount of evaporation which would occur if water was not limited (Hess, 2000) derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005)	480 – 708 mm y^{-1}
AAR	Average annual rainfall derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005)	548 – 1347 mm y^{-1}
RCS	Bedrock geology derived from 1:625,000 scale British Geological Survey rock classification scheme map, detailing bedrock lithology	27
FCD_MED	Median number field capacity days derived from average monthly reports from the UK Meteorological Office on a 5 km x 5 km grid (Perry & Hollis, 2005)	107-290 days
Curvature	Surface curvature derived from a 10m DEM (Childs, 2004)	-74.8 – 66.4
Iwahashi	Iwahashi landform classification uses a terrain classification algorithm based on slope, surface texture and local convexity (Iwahashi & Pike, 2007) derived from a 10m DEM	8
Pennock	Pennock landform classification uses a terrain classification algorithm based on slope, curvature and catchment size (Pennock et al., 1987) derived from a 10m DEM	7

STI	Sediment transport index derived from a 10m DEM	-67.4 - 0
Slope	Slope derived from a 10m DEM (Childs, 2004)	0 – 74.9
SWI	Saga Wetness Index, a terrain-derived index of soil moisture derived from a 10m DEM (Böhner et al., 2001)	9.8 – 19.7
Aspect	Aspect derived from a 10m DEM (Childs, 2004)	-1 - 360

2.2.2 Data Pre-Processing

Models were built using 342 D_b samples from the A Horizon and 339 samples from the subsoil. Many studies differentiate between topsoil and subsoil by depth (De Vos et al., 2005; Katterer et al., 2006). However, the lower depth of the topsoil layer can vary from 15 cm (Bellamy et al., 2005) to 30 cm (Martin et al., 2009). The data used in this study were sampled by horizon, meaning that there was not a uniform sampling depth between points and the number of samples taken at a given location was dependent on soil morphology. As such, the A horizon, with an average depth of just over 22 cm, was used to represent the topsoil. The subsoil layer is comprised of various B horizons (predominantly Bw and Bg) and, on average, represents a horizon between 23 and 47 cm in depth. Of the original samples, the A horizon was split at random into 239 training and 103 validation samples, and the subsoil was split into 238 training and 101 validation samples. Models were built using the training data sampled for each horizon, then these models were applied to the validation data, to provide an unbiased estimate of the predictive power of each model. Descriptive statistics of the soils data are shown in Table 2-3.

Table 2-3: Measured soils data within the study area (A horizon n=342, Subsoil n = 339)

Variable	Mean	Maximum	Minimum	Standard deviation
	A horizon			
Bulk density (g cm^{-3})	1.19	1.76	0.59	0.24
Organic carbon (%)	3.20	15.30	0.50	2.05
Sand (%)	41.33	91.00	3.00	23.35
Silt (%)	35.21	80.00	8.00	15.12
Clay (%)	24.53	74.00	3.00	13.68
	Subsoil			
Bulk density (g cm^{-3})	1.38	1.72	0.80	0.18
Organic carbon (%)	1.03	4.80	0.05	0.73
Sand (%)	42.62	95.10	1.70	25.99
Silt (%)	34.02	79.00	2.90	16.48
Clay (%)	24.35	84.00	0.80	15.81

2.2.3 Statistical methods

In order to develop statistical relationships between a large number of landscape attributes and D_b this study will test both linear (MLR) and non-linear models (RF and ANN). The MLR method is the most frequently used method of modelling D_b, hence these models provide the standard against which the other modelling approaches can be judged. The RF and ANN modelling methods are used to test whether using models which can account for complex, non-linear interactions between variables will improve predictive accuracy. This study will to test two distinct methods; which have previously been successfully applied to the prediction of a range of soil properties including D_b (Tranter et al., 2007), soil texture (Ließ et al., 2012) and NIR spectral reflectance (Rossel & Behrens, 2010). Both non-linear methods are suitable for datasets with

numerous predictors, containing both categorical and continuous data. The reason why RF and ANN models have been selected for this study, as opposed to a host of other data-mining techniques, is that they have been shown to be amongst the most powerful predictive techniques (Prasad et al., 2006; Behrens & Scholten, 2006b; Agyare et al., 2007). It is not the intention of this study to determine the most accurate data-mining technique, but rather to investigate the used of data-mining, hence the models selected are deemed to be appropriate.

2.2.3.1 Multiple Linear Regression

Multiple Linear Regression (MLR) is the most common method of fitting predictive models. The format of MLR models is:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_pX_p + E \qquad (2\text{-}1)$$

where Y is the dependent variable, a is a constant, b_x are coefficients, X_x are predictor variables and E is the error term. The significant variables were chosen by a stepwise selection procedure using the stepAIC function of the MASS package in R (Venables & Ripley, 2002). MLR is not able to explicitly include categorical variable, therefore, to include categorical variables the 'factor' function of R statistical language is used. This is described as 'dummy coding' which effectively stratifies the dataset (Faraway, 2002). As there are limited data, only the Land cover and soil great group are included in the MLR models. Stratification using these variables has been shown to improve the accuracy of regression models for the prediction of D_b (Hollis et al., 2012; Steller et al., 2008). In the subsoil, soils are stratified by parent material rather than soil great group, at this has been shown to improve predictions in deeper horizons (Hallett et al., 1998; Calhoun et al., 2001).

2.2.3.2 Random Forest

RF modelling has the potential to improve predictions made using classification and regression trees (CART) (Breiman, 2001). In essence, trees are constructed using a bootstrap of the entire dataset and the splits at each node are not made by the best predictor from the entire suite of input variables, but from the best of a randomly selected subset, which prevents overfitting (Liaw & Wiener, 2002). The model only requires two user-defined parameters: the number of trees in the forest (n_{tree}) and the number of variables tested at each node (m_{try}). The performance of the training model is assessed by predicting the mean square error (MSE) of the 'out of bag' portion of the data at each tree, then averaging over the entire forest:

$$MSE_{OOB} = n^{-1} \sum_{i=1}^{n} (z_i - \hat{z}_i^{OOB})^2 \qquad (2\text{-}2)$$

where z_i^{OOB} is the mean out of bag prediction for the *i*th observation. RF modelling also provides a measure of fit comparable to the R^2 values of the other models. This 'pseudo R^2' is labeled the 'percent variance explained' and is calculated using:

$$Var_{ex} = 1 - \frac{MSE_{OOB}}{\hat{\sigma}_y^2} \qquad (2\text{-}3)$$

where $\hat{\sigma}_y^2$ is the total variance of the dependent variable calculated with n as the divisor, rather than $n - 1$ (Liaw & Wiener, 2002). The parameters were set to an n_{tree} of 1000 and an m_{try} of $p/3$, where p is the number of input variables. Liaw & Wiener (2002) suggest testing the m_{try} value by both doubling and halving the default. Models can be sensitive to the m_{try} parameter, as testing a greater number of variables at each split will increase the strength of the individual tree but increase the correlation between trees in

the forest. Here the optimal m_{try} was determined using the tuneRF function (Ließ et al., 2012). Furthermore, the n_{tree} value was increased from 500 (the default) to 1000 as recommended by Prasad et al. (2006). This number of trees is sufficiently large to stabilize errors, whilst not being too computationally demanding. An interesting feature of RF is its ability to rank predictor variables in order of importance, which is done by measuring how much the 'out of bag' estimate error increases when data for a particular variable is 'removed' from the analysis and the other variables are left intact. This is done on a tree-by-tree basis for the entire forest. The models were generated using the 'RandomForest' package (Liaw & Wiener, 2002) in the R statistical computing language (RDevelopment Core Team, 2008).

2.2.3.3 Artificial Neural Networks

The principles of ANNs are well established (Bishop, 1995) with Maier & Dandy (2001) offering a practical guide for environmental modelling. The structure used here was a multilayer perceptron, a powerful predictive technique and that most commonly applied in soil science (Agyare et al., 2007). In this method, data are separated into a series of nodes, with similar nodes arranged into layers: typically, an input layer (containing the variables used for prediction), an output layer (where predictions are made) and, in-between, a single hidden layer which weights and transforms the data to extract meaningful relationships (Figure 2-2).

The idea for the ANN modelling approach came from the way data is processed by the human nervous system. Generating predictions from data is a two-stage process. Firstly, the network is trained, which in this instance means that it links predictor variables to D_b values. Using the example in Figure 2-2, each node under the 'Predictor variables' column represents a landscape variable. For example, the nodes P1...P_n represent the

parent material classes present in the training data. The connections between nodes are weights (W_{PV1}... W_{PVn}) which are assigned at random at the start of the training process and are attuned iteratively to best match the training data. To clarify, the network is presented with training data (in this case 239 samples for the A horizon and 238 samples for the subsoil) which it processes individually. The resultant error between the predicted D_b value and the observed D_b values, caused by the weights at each of the links, are recorded and the weightings are subsequently amended to reduce the error. After a number of iterations, the network determines an optimal set of weights, which will be those that minimise the error function (2-4). Secondly, predictions are made using the trained network and the landscape data from unsampled locations (Behrens et al., 2005). The network is structured to include a hidden layer containing a number of nodes (H1... H_n). It is impossible to predetermine the number of optimal number of hidden nodes within the hidden layer, therefore this parameter requires testing (Zhu, 2000). This hidden layer is used to transform or recode the input data, to provide more accurate predictions.

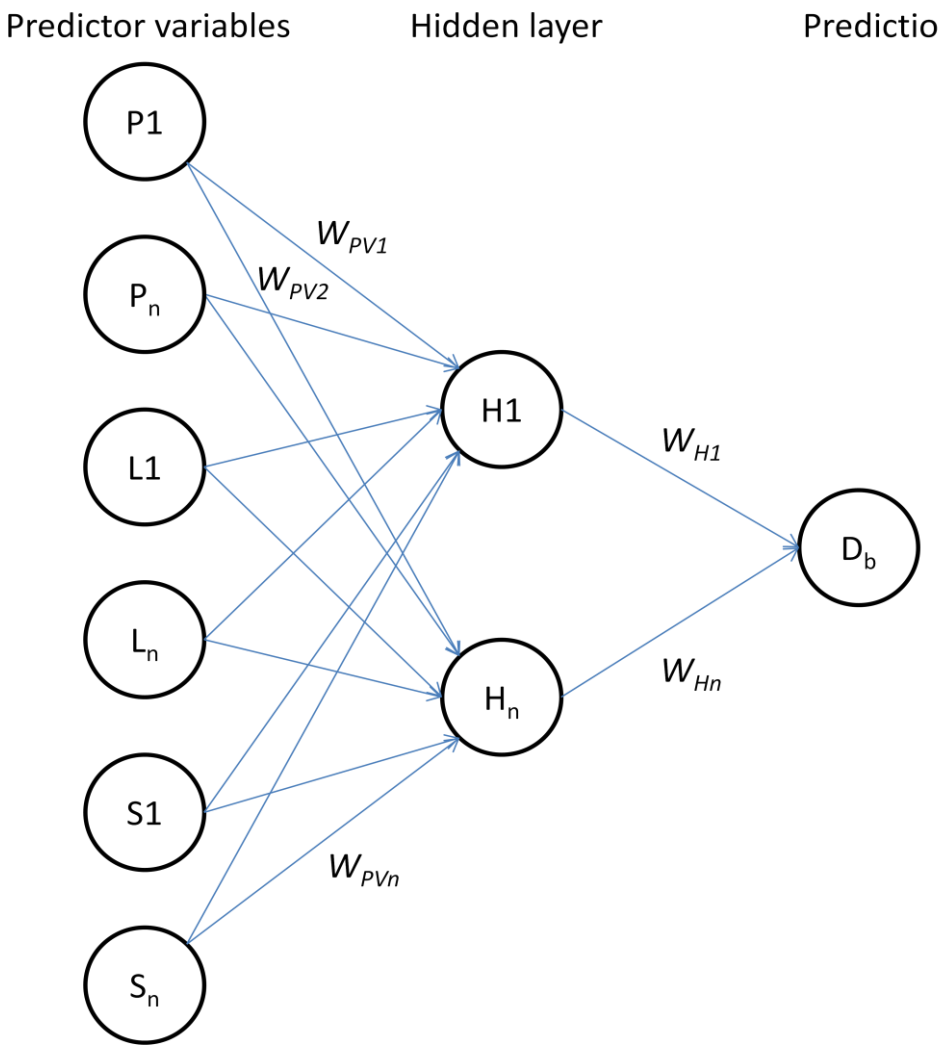


Figure 2-2: Example of the topology of a feed-forward, multilayer neural network (Adapted from Behrens et al., 2005).

For each model, the 239 samples used for developing the models were separated into a 75:25 percent split for training and testing, respectively. As with the other models the remaining 103 samples were used for independent validation. Splitting the data allowed the number of hidden nodes to be tested, which is important as the optimum number of nodes will differ depending on the problem at hand and the number of input variables. It is recommended that the number of hidden nodes should be half the number of input variables plus the number of output variables (which in this case was one) (Statsoft, Inc., 2011). Generally, adding more nodes will increase the performance of the model.

However, this may lead to overfitting the data. To avoid this, the ANN uses a back-propagation algorithm (Rumelhart et al., 1986) to test the performance of the network on both training and testing datasets. Training the network should reduce the 'error function' associated with predictions, such that when the error function of the testing dataset plateaus or increases, ANN overfitting is suggested. The error function for regression is the Sum of Squares error given by:

$$E_{SOS} = \sum_{i-1}^{N} (y_i - t_i)^2 \qquad (2\text{-}4)$$

where N is the number of training cases, y_i is the predicted value of the i^{th} case and t_i is the observed value. Ideally, when the differences in the error function are negligible, the network with the fewest nodes is chosen. As the test dataset plays a role in developing the ANN infrastructure, a validation data set is used to independently test the predictive power of the models. The best performing models were selected using values of R^2 and root-mean-square error (RMSE). ANNs can also rank variables in order of importance, although they use a different method from RFs. Here, data for each variable is replaced, in turn, by its mean value from the training data and the effect on the error function is recorded. The variables are then ranked by the amount their omission increases the overall error function (Lou & Nakai, 2001). The learning rate for the ANNs was set to 0.1 and the analysis was carried out using STATISTICA9 (StatSoft Inc., 2011). One issue arising when using ANNs for producing predictive maps is that they will not make predictions in areas where data differ from those of the training data. In other words, if not every category of, for example, Land cover is included in the training data, the final maps will leave blank areas when they encounter these missing categories as opposed to inferring the D_b values from available data. While this leaves

areas with missing predictions, it means the accuracy of the final map is not compromised.

2.2.3.4 Comparing model performance

The validation dataset was used to accurately compare the predictive capabilities of the three modelling approaches. Each model was used to predict the D_b values in the validation dataset which was then compared to the observed values. The model performance is assessed using the root mean square error given by the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - t_i)^2} \quad (2\text{-}5)$$

Where n is the number of observations, y_i is the predicted value and t_i is the observed value and the coefficient of determination (R^2), given by:

$$R^2 = \frac{\left[cov(y_i, t_i)^2\right]}{var(y_i) \cdot var(t_i)} \quad (2\text{-}6)$$

Where n is the number of observations, y_i is the predicted value and t_i is the observed value, var is the variance and cov is the covariance function.

2.2.4 Calculating OC Stock

To illustrate the importance of D_b for soil inventory, the variation in carbon stock estimation was calculated using measured, predicted and mean D_b values. As a single, unweighted mean across a heterogeneous area would lead to bias results, the mean D_b was calculated for each soil great group (Avery, 1980) and weighted by area. Using a mean D_b value stratified by soil great group is an approach which is commonly used to representing the spatial variation of D_b across the landscape (Grimm et al., 2008; Batjes, 1996). Carbon stock was calculated using:

$$S = d \cdot OC \cdot D_b \cdot 10 \quad (2\text{-}7)$$

where S is the soil organic carbon stock (t C ha^{-1}), d is depth of the topsoil (m), OC is organic carbon concentration per unit mass of dry soil (kg C kg^{-1}) and D_b is soil bulk density (kg m^{-3}). Note that within our calculations, depth is kept constant. To evaluate the uncertainty associated with carbon stock estimation, it is necessary to propagate the errors associated with both OC and D_b measurements and predictions, while keeping depth constant (Schrumpf et al., 2011). The variance is given using the formula:

$$Variance(OC\ Stock) = (OC\ Stock)^2 \cdot \left(\frac{(\sigma OC)^2}{(OC)^2} + \frac{(\sigma D_b)^2}{(D_b)^2} + 2\frac{covOC - D_b}{OC \cdot D_b}\right) \quad (2\text{-}8)$$

where σOC and σD_b are the standard deviations of OC concentration and D_b respectively and $covOC - D_b$ is the covariance between the OC concentration and D_b. In the gridded model, covariance was determined using the predicted D_b values and the measured OC values. In the stratified model, the covariance between the mean great group D_b and OC was determined using a mixed-effects model (Wutzler et al., 2008).

2.3 Results

2.3.1 Model Performance

The results for the MLR, RF and ANN models for both topsoil and subsoil are shown in Table 2-4. For each of the modelling approaches, the A horizon was predicted more accurately than the sub-soil. For the A Horizon, the best performing model was the RF, with the model which combined measure soils data with landscape variables able to describe over 70 percent of the topsoil bulk density. In the subsoil, MLR was the best

performing model, explaining nearly 60 percent of the variation in D_b. It is noteworthy that although the model fit (in terms of R^2 values) is generally worse for the subsoil than for the A Horizon, the RMSE is lower in the subsoil models. This reflects the smaller variation between D_b in subsoil horizons.

Table 2-4: Modelling results (using the validation dataset) for MLR, RF and ANN models. The Suffix 'A' indicates the results are for the A Horizon and the suffix 'S' indicates the results are for the subsoil. For the RF and ANN models, the top five predictor variables are ranked in order of importance, for the MLR model, all variables included in the stepwise models are reported.

Model	R^2	RMSE	Predictor Variables Rank
PTFs (Soil texture and OC content)			
MLR-A	0.4979	0.1627	1. OC 2. Sand
ANN-A	0.6731	0.1306	1. OC 2. Sand 3. Clay
RF-A	0.6665	0.1313	1. OC 2. Sand 3. Clay
MLR-S	0.5131	0.1209	1. OC 2. Sand
ANN-S	0.5529	0.1162	1. OC 2. Sand 3. Clay
RF-S	0.4209	0.1321	1. OC 2. Sand 3. Clay
PTFs and landscape variables			
MLR-A	0.5408	0.1576	OC, Sand, Land cover, Soil group, AT0 annual, PT, Curvature, SWI, Elevation
ANN-A	0.6377	0.1406	1. Land cover 2. Soil Association 3. LEX 4. Iwahashi 5. Parent Material
RF-A	0.7107	0.1323	1. OC 2. Land cover 3. Soil Association 4. Sand 5. Parent Material
MLR-S	0.5901	0.1098	Clay, OC, Parent material, AT0 Annual, FCD Med, Slope
ANN-S	0.4434	0.1310	1. Parent material 2. Land cover 3. LEX 4. Soil Association 5. OC
RF-S	0.5639	0.1163	1. OC 2. Parent Material 3. Soil Association 4. RCS 5. Land cover

Landscape variables only			
MLR-A	0.3692	0.1853	Land cover, soil group, AT0_annual, PT, Curvature, Elevation
ANN-A	0.5507	0.1677	1.Great Group 2.Land cover 3. Bedrock 4.Parent Material 5. FCD_MED
RF-A	0.5602	0.1651	1.Land cover 2. Soil Association 3. Elevation 4. Great group 5. AT0 Annual
MLR-S	0.3016	0.1444	Land cover, Parent material, AT0_annual, FCD Med, Slope
ANN-S	0.3108	0.144	1. Land cover 2. Parent Material 3. Soil Association 4. Bedrock 5. Pennock landscape classification
RF-S	0.2008	0.1581	1. Soil Association 2. Parent material 3. Land cover 4. Bedrock 5. PET

2.3.2 Predictor Variables

Both the RF and ANN modelling approaches have the ability to rank the predictor variables in order of importance. Although they do so in different ways, this means it is possible to assess whether there are common predictors influencing the variation in D_b. the stepwise selection of variables for the MLR models means only variables with improve the predictive power of the models are included. This means it is possible to examine the predictors used in each modelling approach and compare similarities and differences. In the A horizon, the consistently important predictors are Land cover and soil group. Climatic factors also feature as important predictors, with annual average temperature and median field capacity days shown to be significant for the RF and ANN models, respectively. Of the measured soil properties, organic carbon content was a consistently important predictor. The variation in the subsoil layers can be more attributed to a combination of soil association, parent material and bedrock geology.

Table 2-5: Point estimates of OC stock. Average stock was calculated using Equation (2-7). Regarding the prediction methods, 'Measured' uses measured D_b values, 'Gridded' uses the gridded predicted D_b values and 'Mean' uses the measured mean D_b per soil great group.

Prediction method	Average OC stock ($tCha^{-1}$)	Error from measured mean ($tCha^{-1}$) (% in brackets)	5^{th} percentile error ($tCha^{-1}$) (% in brackets)	95^{th} percentile error ($tCha^{-1}$) (% in brackets)
Measured	73.01±0.56	NA	NA	NA
Gridded	71.32±0.61	1.69 (-2.31%)	5.71 (-15.43%)	10.79 (8.37%)
Great Group mean	74.81±0.70	1.80 (2.47%)	6.34 (-17.14%)	19.31 (14.99%)

Table 2-6: Carbon stock for the entire study area and by selected Soilscape

Location	OC Stock ($t\ ha^{-1}$) estimated using great group mean D_b (± 95% confidence interval)	OC Stock ($t\ ha^{-1}$) estimated using gridded D_b (± 95% confidence interval)
Full study area	86.41±15.59	87.01 ± 8.19
Central England Plateau	84.72 ± 15.01	88.25 ± 8.18
Central upland spine of N England	86.75 ± 16.98	71.84± 8.41
Total Carbon Inventory (Tonnes)	156834150 ± 28295850	157923150 ± 14862371

2.4 Discussion

2.4.1 Model Performance

Random Forests were able to describe D_b most effectively, which is unsurprising as they are designed specifically for large, heterogeneous datasets containing a mixture of both continuous and categorical variables (Liaw & Wiener, 2002). Indeed, tree-based

models have been used to successfully predict D_b using a mix of landscape data and soils data (Martin et al., 2009). In terms of model performance, RF achieved better results than a number of comparable studies (Tranter et al., 2007; Martin et al., 2009; Hollis et al., 2012). The ANN model also performed well for the A horizon. Previous studies (e.g. Minasny et al., 1999, Keshavarzi et al., 2010) have reported both high and low ANN performance. This can be attributed to the nature of the property being predicted. Wösten et al. (2001) suggest that generally, when there are more than three predictor variables and variables are subject to complex interactions, non-linear modelling techniques such as AAN and RF become necessary. In this instance, for the A horizon at least, it appears that the prediction of D_b can be improved through the use of non-linear modelling approaches, even if the predictor variables are limited to OC content and textural properties. The decrease in predictive power in the ANN models when the landscape variables were included can be attributed to the inclusion of variables which are not strongly correlated to D_b. According to Behrens et al. (2005), the inclusion of extraneous variables should not negatively impact on the performance of an ANN model, however, other studies suggest that given this scenario, ANNs are prone to overfitting (Amini et al., 2005). The subsoil was generally less well predicted, although the relatively high predictive power of the MLR model suggests that stratification by parent material is sufficient to describe the variation of D_b in the subsoil horizons using a linear model. This reflects the findings of Hallett et al. (1998) who demonstrated that stratification by lithology produced more accurate predictions of D_b in the subsoil, compared to stratification by soil group. The poor performance of both the RF and ANN models in the subsoil layer reflects the lower spatial variability of the

subsoil D_b (Braakhekke et al., 2013), meaning changes in landscape predictors exhibit relatively little influence.

2.4.2 Variable Importance

It has been well established that OC content is usually the most important predictor when modelling D_b. This is unsurprising as the relationship between the two has been well defined (Rawls, 1983) and used extensively in predictive modelling (Kaur et al., 2002). However, Calhoun et al. (2001) found that particle size distribution and OC generally explain no more than 60 percent of the variation in bulk density. Of particular interest here is the predictive power of the seldom-used variables which represent a range of topographic, Land cover and climatic factors. The importance of putting D_b in a landscape context is supported by the successful stratification of previous regression models by Land cover (Steller et al., 2008; Moreira et al., 2009) and parent material (Hallett et al., 1998; Calhoun et al., 2001). However, these factors have been explicitly included in the modelling process only relatively recently (Martin et al., 2009; Jalabert et al., 2010). Of the landscape variables included, Land cover, parent material and soil classification are deemed to be consistently important predictors. The influence of soil class is unsurprising as, along with other attributes, soils are classified based on their textural properties. Using pre-existing soil maps is, in essence, a way of predicting using spatially distributed textural classes. The predictive power of Land cover will depend on the classification used and the resolution of the data layer. Previous prediction of D_b using boosted regression trees by Buttner et al. (2000) has suggested that Land cover derived from the European CORINE map was the least influential of all their predictor variables, as these Land cover classes were too broad. However, more detailed, higher resolution Land cover information transpired to be the second most

powerful explanatory variable, almost on a par with OC content (Jalabert et al., 2010). As Land cover was recorded at the time of sampling, the accuracy of the layer was not in question, and hence it proved to be an important predictor. To make use of use of the available Land cover data, the CEH Land Cover Map was re-coded to reflect the Land cover at the time of sampling. This was important as, when used as a predictor without re-coding, present day Land cover categories were shown to be poor predictors of D_b. This can probably be attributed to the fact that sampling of D_b and the creation of the Land cover layer were approximately 30 years apart, with significant changes over the intervening decades.

Parent material is one of the leading predictors in nine of the twelve models in which it is included. This may be attributed to the presence of recently deposited material, such as alluvium, or slow draining or impermeable bedrock which are particularly influential for overlying soil formation (Hallett et al., 1998). Pertinently, a significant number of samples in this study were taken from alluvial plains, in which soil properties, such as D_b, are closely related to the properties of the underlying alluvium, thereby promoting the influence of parent material as a significant predictor. In other areas with less alluvium, parent material may be less influential on D_b. Predictably, parent material becomes a more influential predictor in subsoil horizons, which are less susceptible to climatic changes. Bedrock geology also becomes more influential below the A horizon.

It is interesting that the climatic variables are such prominent predictors because they have a relatively low spatial resolution (5 km grid), in comparison with other predictor variables, although the link with some variables (e.g. field capacity) has clear physical significance. This suggests that improving the resolution of climatic predictors may improve model accuracy. The DTM-derived landscape attributes proved to be relatively

poor predictors. Although Martin et al. (2011) mention including topographic predictors as a possible improvement for mapping OC stocks, they are not generally utilized. In similar work to model saturated hydraulic conductivity, landscape derivatives have offered some improvement to ANN models, but they cannot be used without other inputs – particularly at a regional scale (Agyare et al., 2007), this reflects the inclusion of elevation as a prominent predictor in the landscape-only RF model.

2.4.3 Modelling without using measured soil properties

Mapping D_b without point samples of soil properties is of interest for two reasons. Firstly, since the cost of large scale soil sampling can be prohibitive, the ability to use pre-existing or remotely sensed data would be desirable. As many countries already have soil, Land cover and geological maps at a variety of scales, it makes sense to see if further information can be extracted from them in the form of predictive models. Secondly, a key research theme in spatial mapping is the assessment soil carbon stocks because they relate to the global carbon budget (Bellamy et al., 2005; Tornquist et al., 2009; Wiesmeier, et al., 2011). One issue regarding the derivation of soil carbon stocks is the lack of spatial representations of D_b. Instead, mean D_b values are used to convert modeled SOC concentrations into SOC stocks (Grimm et al., 2008). However, if variations in D_b within individual soil types are not taken into account, significant errors in C stock estimation are possible. As datasets tend to be limited, and OC and D_b are not always sampled together, being able to map D_b accurately and independently of measured OC content, would avoid circularity in modelling (i.e. using carbon content to predict D_b which is then used to predict carbon stocks) and improve stock estimation at the same time. This study has determined that many of the important predictor variables are categorical (Land cover, parent material) and using landscape variables alone, for

the A horizon, both RF and ANN techniques can explain over 55 percent of the variation in D_b. This result is significant because it shows that it is feasible to create a continuous surface of D_b solely using landscape attributes. A spatial representation of D_b across the landscape can be combined with a spatial representation of carbon concentration to give a more accurate estimate of C stocks and pools. At any given location, there will be an associated D_b value, at an appropriate scale, which has been independently derived and which has an associated unambiguous error estimate.

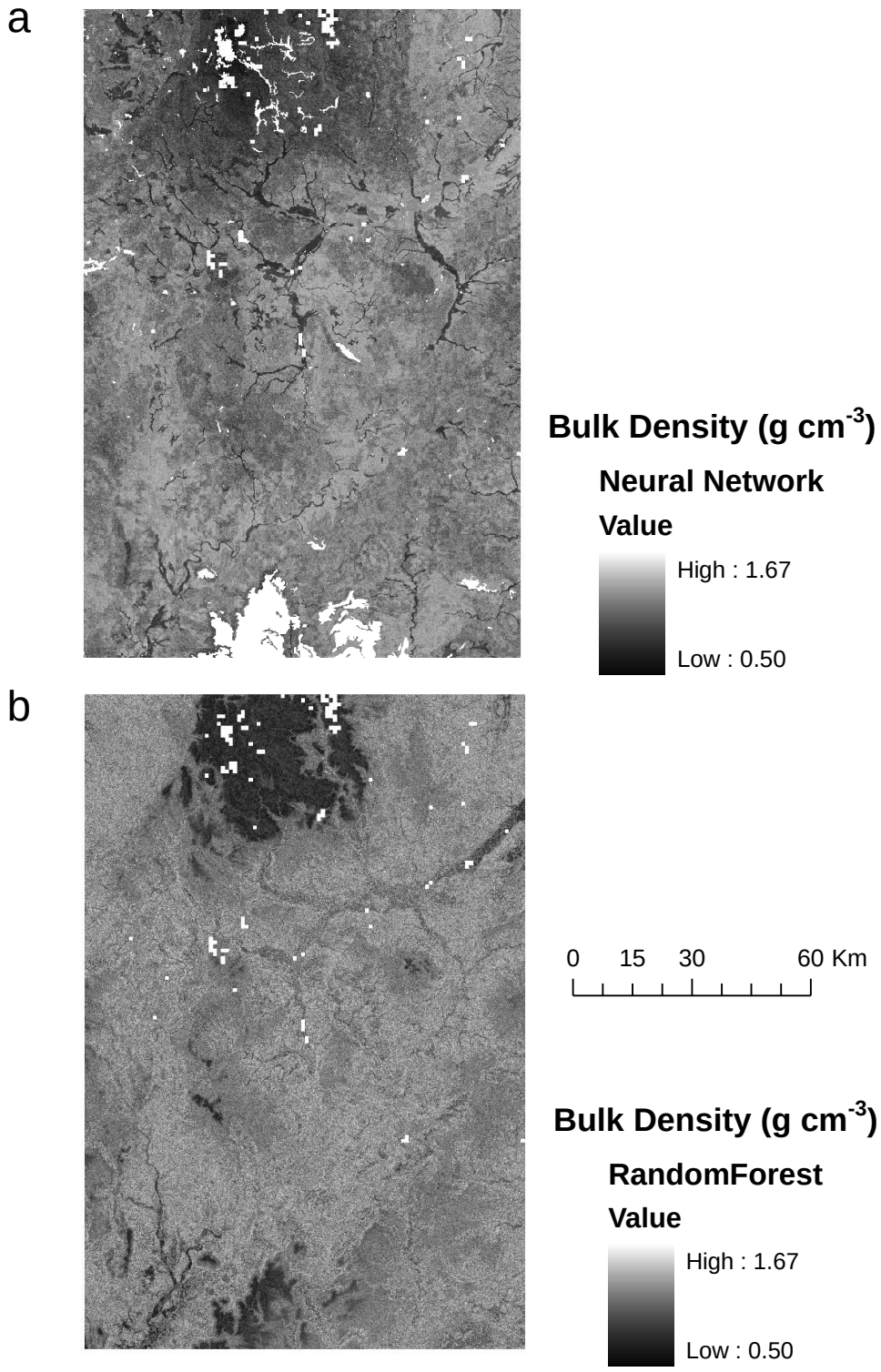


Figure 2-3: Predicted bulk density across the landscape obtained from models built using the training dataset. a) Artificial neural network b) Random forest.

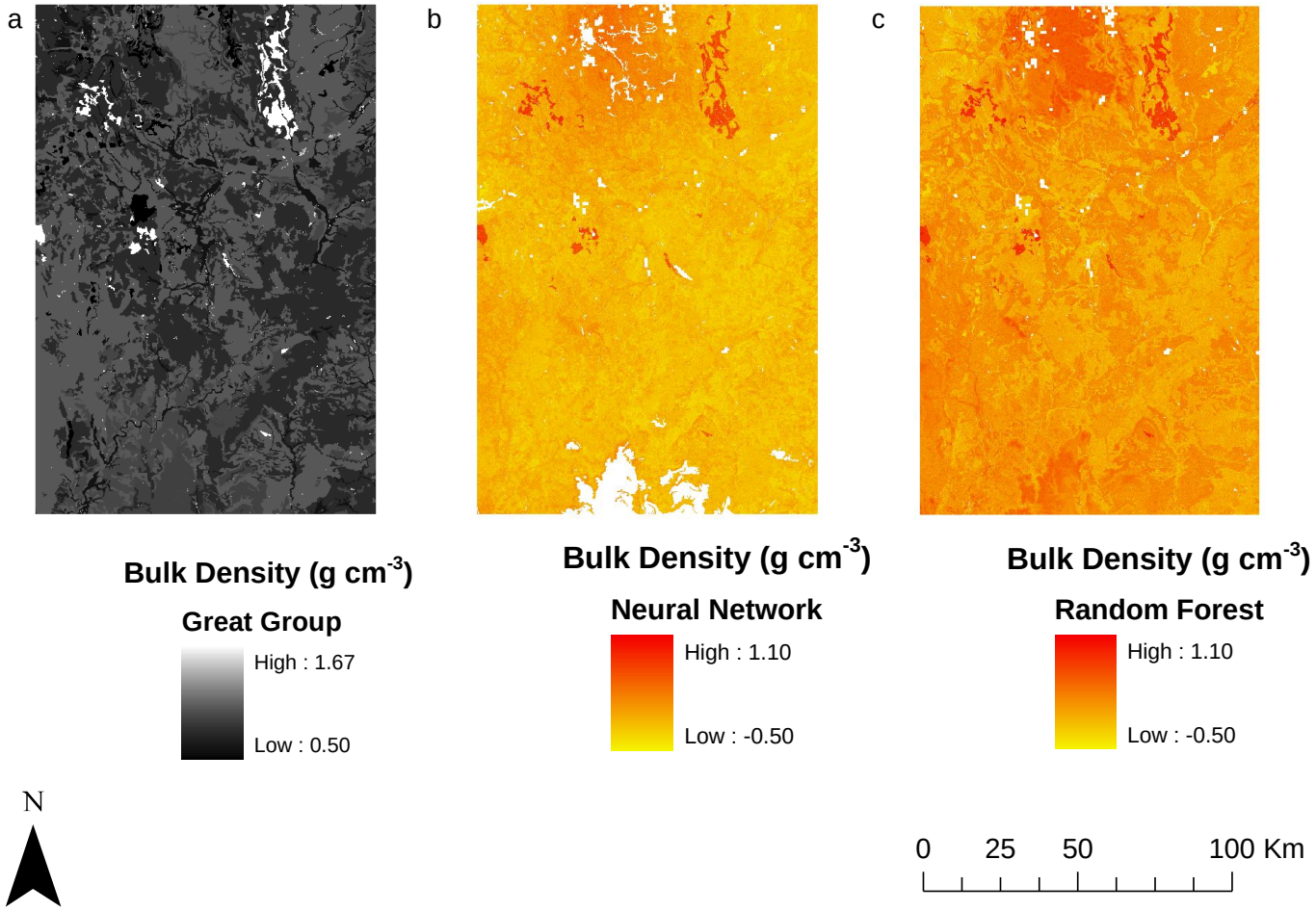


Figure 2-4: Difference map of bulk density predictions. a) Great Group bulk density b) Difference in Neural Network prediction c) Difference in Random Forest prediction

2.4.4 Mapping D_b across the landscape

For the A horizon, maps of D_b for the topsoil of the entire study area have been produced using both ANN and RF (Figure 2-3). Topsoil is generally considered to be the most important soil compartment in terms of soil carbon content, in part because OC concentration generally decreases with depth (Jones et al., 2005). Of the two methods, ANN gives a slightly wider range of predicted D_b values than RF but still within the limits of the measured data reported within the National Soil Inventory of England and Wales (Loveland, 1990). Fewer than three percent of the samples in the National Soil Inventory had a D_b lower than the minimum predicted value. In contrast, RF (Figure 2-3b) provides more conservative estimates of D_b, especially for the upper values. Despite this, the RF model was shown to have slightly more predictive power than the ANN model. Broadly speaking, the models agree on the spatial trends of D_b distribution, most notably, areas of low D_b in the north and at the westerly edge of the study area. The areas of missing data in the ANN model reflect missing data in the training dataset. Here the RF models are used to make predictions based on the available data. The difference map (Figure 2-4) is used to illustrate areas where the RF and NN model predictions differ. Figure 2-4a show the D_b averaged by soil great group and the subsequent two figures (2-4b Neural Network and 2-4c Random Forest, respectively) show how the predictions differ from this baseline map. Generally, there is good spatial agreement between the modelling techniques, although, noticeably, RF modelling appears to predict lower D_b values across the region.

2.4.5 Spatial Performance

Spatially, there is broad agreement between the RF and ANN predictions, in terms of the areas of high and low D_b. Figure 2-5 shows the individual performance of each

model, in terms of prediction residuals as an average per Soilscape. In the A horizon, the spatial variation in the relative performance of each statistical approach is very similar (Figure 2-5b and Figure 2-5c). In terms of Land cover and soil group, the two most influential predictors of topsoil D_b, both the RF and ANN models give their best predictions in areas of Brown Earths under arable Land cover. The areas across which both models appear to perform least well coincide with built up areas dominated with Stagnogley soils. In the subsoil, the spatial patterns of model performance are also broadly similar for both the ANN and RF models. In relation to parent material, the best predicted regions coincide with areas of sandstone bedrock and superficial deposits containing siliceous stones while the worst performing areas overly clay or soft mudstone. The spatial variation in model performance can be used to inform any future sampling schemes, with an increased sample density in areas where a model is likely to underperform.

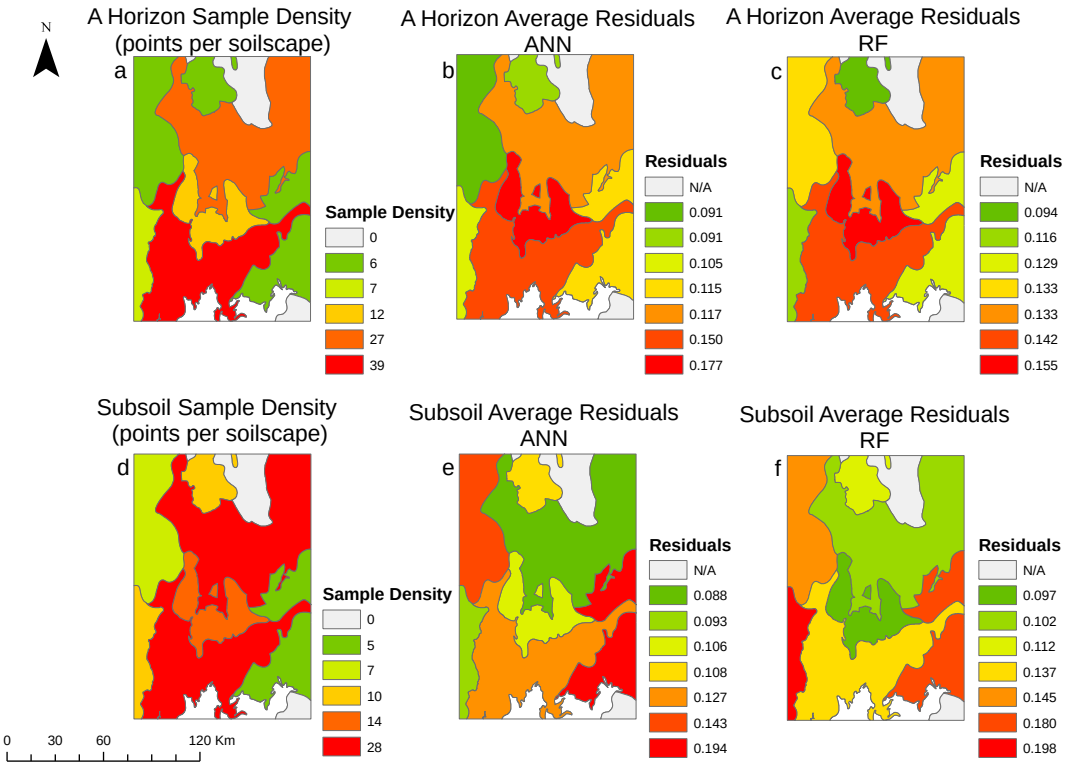


Figure 2-5: Spatial variation in model performance by Soilscape. a) The sample density for A horizon samples b) Average residuals for the ANN model prediction in the A horizon c) Average residuals for the RF model prediction in the A horizon d) The sample density for subsoil horizon samples e) Average residuals for the ANN model prediction in the subsoil horizon f) Average residuals for the RF model prediction in the subsoil horizon

2.4.6 Stock Estimation

To illustrate the potential improvement in OC stock estimation which could be achieved using the gridded surface of D_b compared with using a stratified mean value (Mestdagh et al., 2009; Hanegraaf et al., 2009) the OC stock at each sample point was calculated using three different sets of D_b: the measured D_b, the RF gridded prediction of D_b and great group mean measured value of D_b calculated using all sample points in the training data. Note that results for C stock calculations using model output were produced using a calibrated RF model that used the training dataset alone, the validation

data was used solely to assess model performance. The average OC stocks calculated using each D_b estimate are shown in , along with the difference between the estimated and measured mean OC value, expressed as a percentage of the mean measured value. The 5^{th} and 95^{th} percentile errors in measured OC stocks are also shown. The gridded surface refers to a map of RF-predicted D_b values (Figure 2-3b) produced as a raster grid with a cell size of 100 x 100 m across the entire study area. The main advantage of the gridded surface method over PTFs, which can be applied to individual points using measured soil property data for the point in question, is that the gridded method can be applied to the entire study area with the same quantifiable level of both performance and error estimation at all spatial locations. In contrast, the accuracy of predictions made using a PTF is hard to quantify beyond each sampling point.

Using the individual measured point-based D_b values gives an average OC content of 73.01 ± 0.56 t C ha^{-1} compared to an average value of 71.32 ± 0.61 t C ha^{-1} produced using the RF-predicted D_b values and a value of 74.81 ± 0.70 t C ha^{-1} generated using Great Group mean D_b value. Using the OC stock calculated with measured D_b as a yardstick, the gridded estimate of D_b yields a marginally better C stock estimate compared with using a single (mean) D_b value. In this case, the RF predictions will underestimate D_b whereas using a stratified mean value will overestimate. The difference in the error associated with stock prediction using the gridded D_b values compared to using the mean value of D_b is particularly evident when predicting C stock levels in soils at the extremes of the expected range (i.e. the prediction errors for the 5^{th} and 95^{th} percentile OC stock values). The potential improvement in using the gridded estimate of D_b is most evident in the 95^{th} percentile, where using a stratified mean D_b value will yield an error nearly two times larger.

To put the magnitude of the errors illustrated in Table 2-5 into context, Bellamy et al. (2005) suggest that the average annual rate of change in the OC content for UK topsoil is 0.67g $kg^{-1}yr^{-1}$, which equates to approximately 1.79 t C ha^{-1} yr^{-1}. As the rate of change is comparable in magnitude to the error associated with prediction, it is clearly important to keep error to a minimum if stock changes are to be quantified accurately. The total soil OC inventory across the whole study area, calculated using both the stratified mean and gridded D_b estimates, is shown in Table 2-6. There is a slight difference in the OC stock per unit area (0.6 t ha^{-1}) which equates to a difference of over one million tonnes of carbon for this study area alone. The most notable difference between the stratified mean and gridded approaches to D_b prediction is the error associated with prediction. The 95% confidence interval associated with the stratified mean model is nearly twice as large as that of the gridded model. When estimating the total C stock within the study area, this translates to a difference of over 13 million t C^{-1}.

To further illustrate the potential of this method, carbon stocks were calculated for the landscape as a whole and for two selected individual Soilscapes using both the great group measured mean and gridded predictions of D_b. Soilscapes were selected to represent the range of D_b values within the study area. Results are shown in Table 2-6. The two Soilscapes; the Central Upland Spine of Northern England and the Central England Plateau show areas of relatively low and high D_b, respectively. These regional differences in stock calculations, particularly in the Central Upland Spine of Northern England, highlight potential errors which can be introduced to a stock calculation by using a mean D_b value, depending on the scale of the study. Moreover, the gridded model has a much greater predictive accuracy, with confidence bounds nearly two times smaller compared to the stratified mean model. The mean model produces similar stock

predictions for both the entire study area and the selected Soilscapes. This is a problem as, at the Soilscape scale, the stratified mean model may either under or overestimate carbon stocks (as it appears to have for C stocks in the 'Central Upland Spine of Northern England' Soilscape). This issue does not affect the gridded model, because it is able to apply rules learned across the entire study region, to identify areas of high and low bulk density, a key advantage when working at this scale. A scale at which errors in D_b estimation have shown to be highly significant to carbon stock inventory (Goidts et al., 2009). Estimating C stocks and changes, especially at finer spatial scales requires the use of refined estimates of Db, which can be obtained using the types of landscape-scale models described in this paper. It is at these scales that many spatially distributed land-atmosphere interaction models such as JULES operate (Harrison et al., 2008).

2.5 Conclusions

For the A horizon, using non-paramedic, non-linear models to predict soil D_b will improve predictive accuracy, even if the soil textural properties and OC content are the only predictors used in the model. These predictions can be further improved by the inclusion of landscape variables, however, careful consideration should be given as to which variables are included. This is especially true if using an ANN model, where predictive accuracy decreased with the addition of landscape variables. This study found that it is possible to predict soil D_b solely using landscape derivatives, such as Land cover, geology and climatic data, if only for the topsoil. In this case, of the three statistical modelling techniques tested, RF marginally provided the best results for the A horizon, while ANN performed best for the subsoil. In comparison to previous studies, which have attempted to predict D_b from soil property data, the models constructed in this study were able to provide similar results, in terms of model performance, without

using soil texture or OC content as predictors. The suite of landscape derivatives used was able to explain over 55 percent of the variation in topsoil D_b.

The advantage of this approach is the models' potential to improve the accuracy of other models, in this case soil carbon stock estimates at a landscape scale. Predicting D_b without using point-scale measurements as explanatory variables means that it is possible to create a continuous, gridded surface of D_b without interpolation which can be used in combination with continuous surfaces of predicted soil carbon content to improve estimations of carbon stock. In addition, the technique yields a more accurate measurement of the error associated with such predictions. In terms of carbon stock prediction, the gridded D_b estimate offers a significant improvement in accuracy compared with using a stratified mean value of D_b. In particular, this approach is valuable when applied at a sub-landscape, regional scale, especially in data-poor areas.

3 Using Bayesian Networks for Digital Soil Mapping

Two corresponding issues concerning digital soil mapping (DSM) are the demand for up-to-date, fine resolution soils data and the need to determine soil-landscape relationships. This chapter proposes that a Bayesian network framework is a suitable modelling approach to fulfil these requirements. Bayesian networks are graphical probabilistic models in which predictions are obtained using prior probabilities derived from either measured data or expert opinion. They represent cause and effect relationships through connections in a network system. The advantage of the Bayesian networks approach is that the models are easy to interpret and the uncertainty inherent in the relationships between variables can be expressed in terms of probability. This chapter will define the fundamentals of a Bayesian network and the probability theory which underpins predictions. The study will then demonstrate how Bayesian networks can be applied to the prediction of soil properties, in this case, soil bulk density.

3.1 Introduction

To satisfy the growing demand for up-to-date, fine resolution soils data, there is a call to fully explore the potential of current mapping and modelling software, and apply existing modelling techniques in novel and innovative ways (Hartemink & McBratney, 2008). Predictive modelling of the spatial pattern of soil types and properties is based on a quasi-mechanistic understanding of soil formation and the factors which drive soil variation in the landscape, namely the ClORPT factors (Climate, Organic activity, Relief, Parent material and Time; Jenny, 1941). The relationships between soil forming factors and soil properties are complex and several non-linear modelling techniques have been employed to represent them including Random Forests (Liaw & Wiener, 2002, Grimm et al., 2008, Wiesmeier et al., 2011) and Artificial Neural Networks

(Agyare et al., 2007, Zhao, et al., 2010). A principal disadvantage of these methods is that they are 'black-box', meaning that it is often difficult to interpret the relationship between response and predictor variables in physical terms (Suuster et al., 2012). In Bayesian networks (BNs) the relationship between soil forming factors and soil properties can be directly addressed (Tavares Wahren et al., 2012). Many significant soil processes, such as the terrestrial carbon cycle, are not particularly well understood at the landscape scale and would benefit from the clarity and insight provided by BN modelling (e.g. Braakhekke et al., 2013). Chen & Pollino (2012) state that improving system understanding is a key motivation for using a BN.

Bayesian Networks (BNs) are graphical probabilistic models in which predictions are obtained using prior probabilities derived from either measured data or expert opinion. They represent cause and effect relationships via connections in a network system (Hough et al., 2010) but they differ from other network based methods, such as ANNs, in that the structure of the network and the interactions between nodes are defined by the user based on prevailing process understanding. To clarify, the network structure can be used to represent current understanding of how variables interact with one another at the scale which the study is being conducted. BNs are a flexible way of structuring process understanding stochastically and, unlike purely deterministic models, reflect the uncertainty surrounding cause-effect relationships (one event leading to another) by expressing each relationship as a probability (Dlamini, 2011). They are also ideal for addressing problems where data are limited (Kuhnert & Hayes, 2009). They are frequently applied to ecological systems (McCann et al., 2006), notably conservation (McCloskey et al., 2011), habitat mapping (Smith et al., 2007) and risk mapping of events such as wildfire (Dlamini, 2011) and peat erosion (Aalders et al.,

2011). Bayesian modelling approaches have previously been applied to modelling class (Skidmore et al., 1996; Bui et al., 1999) or soil attribute (Corner et al., 2002), with Bayesian networks, specifically used to predict soil class in the UK (Mayr et al, 2008; Mayr & Palmer, 2006). Despite this, the potential for predicting the distribution of soil attributes and classes using Bayesian networks is yet to be fully explored.

Bayesian networks were developed from the branch of mathematics known as probability theory, in particular from probabilistic reasoning (Pearl, 1988). Unlike deterministic models, BNs offer a structured method of dealing with uncertainty which, as a rule, diminishes as more information is gathered. In the case of predicting the spatial distribution of soil classes and properties, the relationships between variables are highly uncertain and data availability is often limited, so BN's have great potential as a predictive tool (Finke, 2012). Another appealing aspect of BNs is their ability to integrate expert knowledge into the model which can be used to supplement measured data, or define relationships between variables directly. There has been a long-standing drive to formally introduce expert knowledge into soil mapping, usually focusing on fuzzy set theory or possibility theory (McBratney & Odeh, 1997). In contrast, BNs use probability theory, which can be seen to offer a more coherent structure to decision making problems (Degroot, 1988), although, there has been some debate as to which is the superior approach (Krueger et al., 2012).

3.1.1 Theory

BNs are named after the Reverend Thomas Bayes who, in the 18th century, developed a theorem regarding changing probabilities given new information (Bayes, 1783). The basis of a BN is conditional probability, which can be explained using an example from Jensen (1996), where a statement of conditional probability reads

"Given an event B, the probability of event A is x"

In mathematical notation this would read

$$P(A|B) = x \tag{3-1}$$

This statement holds true, only if all other information which could affect event A is known and has been accounted for. The basic rule of conditional probability is:

$$P(A|B)P(B) = P(A,B) \tag{3-2}$$

Where $P(A,B)$ is the probability of the joint event A and B both being true ($A \wedge B$). From this, the Bayes Rule (3-3) can be derived.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3-3}$$

This rule forms the basis of BN modelling, as Bayes' rule is used to inform us of the probability of event A given information about B. Referring to Equation (3-1), the posterior probability $P(A|B)$ was an unknown x, now it can be calculated using our prior belief in the occurrence of event A $P(A)$ and event B $P(B)$ and the probability that B will happen if A is true $P(B|A)$. This is known as Bayesian inference and to illustrate how this might work in practice for digital soil mapping applications, an example given by Aitkenhead & Aalders (2009) has been adapted.

From Equation (**3-3**), $P(A|B)$ is the posterior probability of event A (e.g. high bulk density; D_b) given B (e.g. arable Land cover) (note that the class 'high bulk density' is an example of discretization of a continuous variable into a set of classes, the boundaries of which would need to be defined). $P(A)$ is the probability that bulk density is 'large' (a prior probability derived from either data i.e. the percentage of samples

recorded as high or from expert opinion), $P(B)$ is the probability of the occurrence of arable land (proportion of the study area that is arable land) and $P(B|A)$ is the prior probability that high bulk density samples will be taken from arable land. Let us assume that of the total number of D_b samples, 30% are classed as large, i.e. $P(A)$ = 0.3. In addition let us assume that, 40% of the terrain in the study area is classed as arable, i.e. $P(B)$ = 0.4., and the proportion of high D_b samples found on arable land is 50% i.e. prior probability $P(B|A)$ = 0.5. This probability can be generated either by expert knowledge or using observed data. Combined, these probabilities give the posterior probability that if the land is arable, the bulk density will be high, $P(A|B)$. In this instance

$$P(A|B) = \frac{0.5 * 0.3}{0.4} = 0.375$$

There is a 37.5% probability that D_b will be high on arable land.

In reality, when dealing with complex problems in soil mapping, there will be numerous factors which influence variables of interest. Hence BNs are designed to link large numbers of influencing variables and combine the conditional probabilities of each. BNs comprise two components; 1) a directed acyclic graph (DAG), where each node represents a variable in which the directed links between nodes represents the conditional dependencies of the model and 2) a quantitative component of a network consisting of conditional probability tables (CPT) that accompany each node, which define the dependencies of each variable. Each CPT contains a list of possible states which could be applied to the variable. Using an example adapted from Nadkarni &

Shenoy (2004), Figure 3-1 shows a Bayesian network comprised of four variables: Land cover (L), Soil group (S), Organic carbon content (C) and Soil bulk density (D). The directional arrows between variables indicate causality. The variables with arrows leading into them are known as the 'child nodes' and the variables where the arrows originate are known as 'parent nodes'. Each state is mutually exclusive and the list is definitive; for clarity, the number of states in Figure 3-1 have been kept to a minimum.

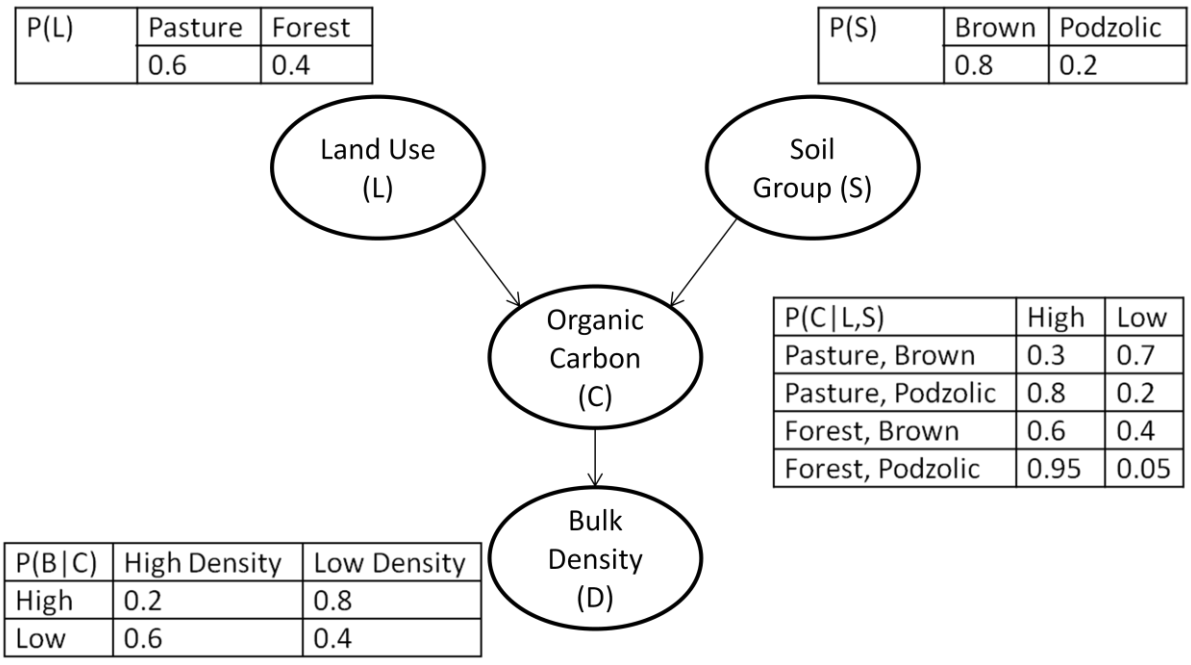


Figure 3-1: An example Bayesian network of soil properties and influencing factors (adapted from Nadkarni & Shenoy, 2004), showing the conditional probability tables for each node

There are three types of connection in a BN (Jensen, 1996). In a serial connection (Figure 3-2) evidence about Land cover (L) is transmitted through C to D. If node C is known (there is hard evidence about Organic carbon), knowledge about Land cover (L) does not transmit to Bulk Density (D), hence any new evidence about L will not change our belief about D. This is known as D-separation where L and D are d-separated given C.

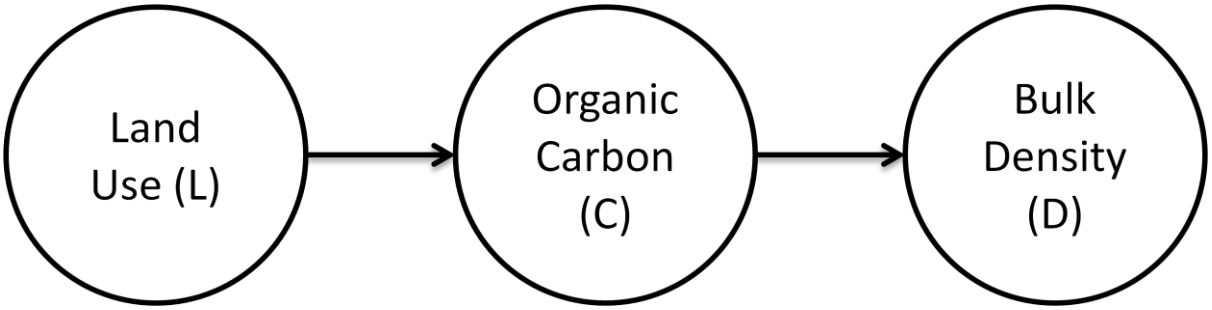


Figure 3-2: An example of a serial connection

In a diverging connection (Figure 3-3), evidence about Parent material (P) is transmitted to both Land cover (L) and Soil group (S). If there is no hard evidence about the state of P, evidence about L can be transmitted to S. When there is hard evidence about P (Parent material is referred to as being 'instantiated'), evidence about L does not transmit to S. When P is certain, L and S become (conditionally) independent, hence L and S are d-separated given P

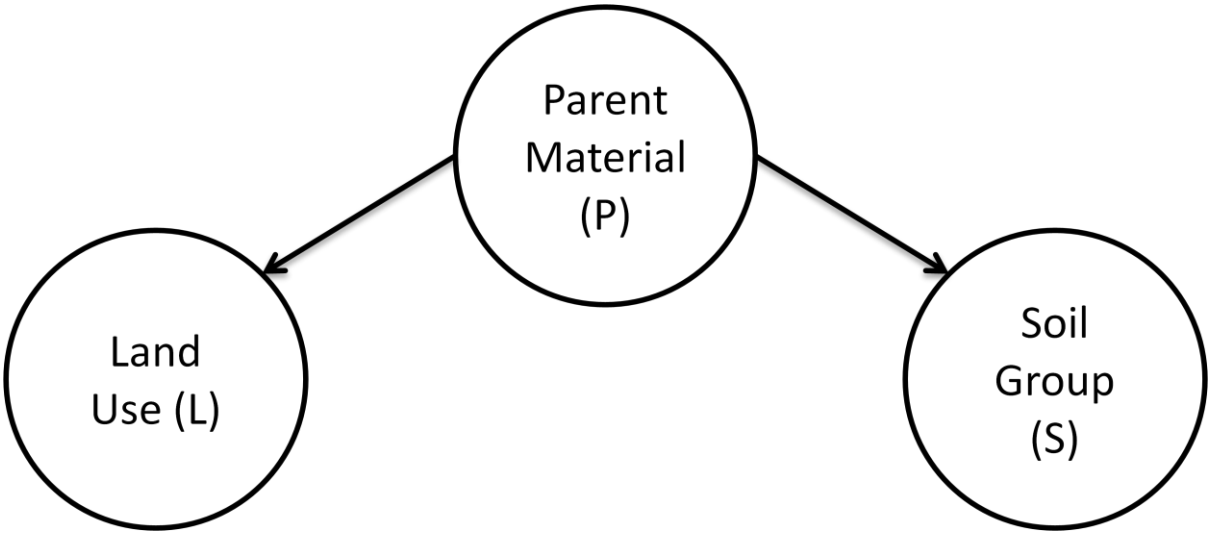


Figure 3-3: An example of a diverging connection

The third type of connection is a converging connection (Figure 3-4). Here, evidence about Land cover (L) and Soil group (S) is transmitted to Organic carbon (C). If nothing is known about C than L and S are independent (information is not passed between

them). This is important as often BNs will be used to determine the most likely cause of an event given evidence. However, if anything is known about C (including 'soft evidence', which may not determine the state of C, but may alter its probability distribution), then L and S become (conditionally) dependent.

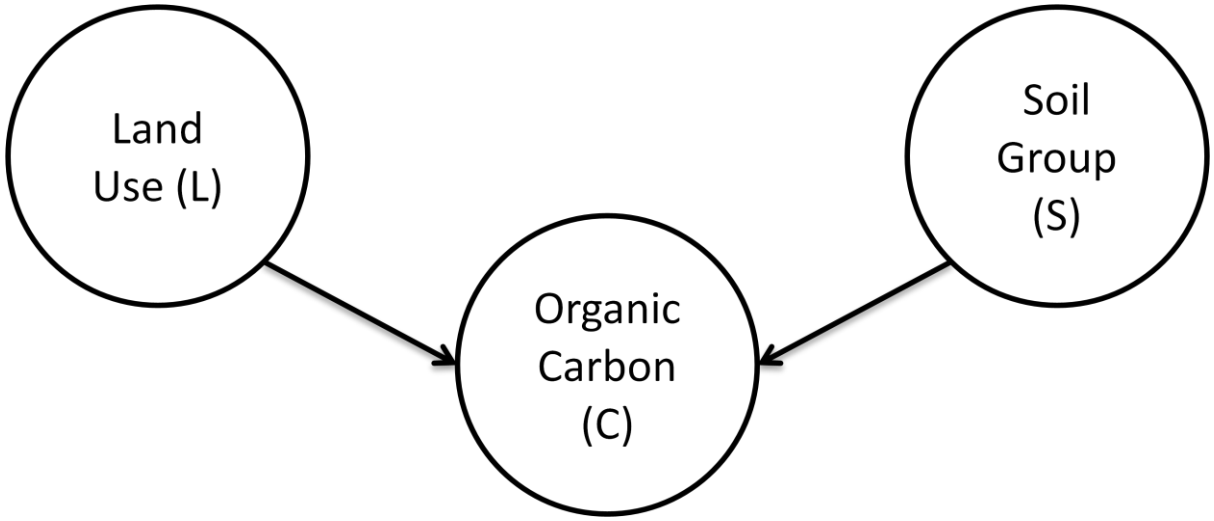


Figure 3-4: An example of a converging connection

In order to function, BNs rely on certain independence assumptions. The links between nodes indicate what information about probabilities is required to produce the probability distribution at the node of interest. All parentless nodes need to be supplied with a prior probability and all child nodes need to have a conditional probability table of every combination of the parent node. Expanding the fundamental rule of conditional probability (Equation (**3-2**)) to incorporate n variables provides the chain rule (Equation (3-4)), which allows us to calculate the full joint probability for all the variables in the network.

$$P(A1, A2, \dots, A_n) = P(A1|A2, \dots, A_n)P(A2|A3, \dots, A_n)P(A_n - 1|A_n)P(A_n) \qquad (3\text{-}4)$$

There is, however, a practical drawback with this rule in its current form. For example, with n random binary variables, the number of joint probabilities required is$2^n - 1$.

This can quickly escalate to a very large number when a network represents the natural environment due to the large number of potentially interacting variables and the fact that variables will often have more than just two states (i.e. they will not be binary). In order to avoid this issue, BNs use an assumption of independence, which reduces the number of probabilities which need to be specified (Charniak, 1991). This assumption relates to how evidence is transmitted through the network and how the probabilistic relationship between variables within the network can be interpreted depending on how they are linked. This assumption is clearly not correct in natural complex systems, as in reality, very few environmental covariates will be completely independent from one another. All modelling approaches will make assumptions in order to represent the natural environment; in a BN these assumptions are explicit (as opposed to the unknown assumptions used in data-mining methods). Once the network is complete and has been tested, it is possible to examine and change these assumptions in order to better represent the system being modelled.

Given these assumptions about conditional probability, it is possible to re-write Equation (3-4) as:

$$P(A_1, \dots A_n) = \prod_{i-1}^{n} P(A_i \,|parents(A_i)) \tag{3-5}$$

where $parents(A_i)$ is the set of parent nodes for variable A_i. In this way, the joint probability distribution of the nodes in a BN is greatly simplified. Using the network in Figure 3-1 as an example, without the assumption of conditional independence, the joint probability for the network is:

$$P(L, S, C, D) = P(L)P(S|L)P(C|L, S)P(D|L, S, C) \tag{3-6}$$

where L denotes Land cover, S denotes Soil group, C denotes Organic carbon, and D denotes Bulk density. However, by assuming the nodes are conditionally independent, the joint probability is given by:

$$P(L,S,C,D) = P(L)P(S)P(C|L,S)P(D|C) \tag{3-7}$$

Note that here it is assumed that $P(S) = P(S|L)$, meaning that the probability of event S is the same as the probability of event S given L, making S independent of L. A further assumption is that $P(D|C) = P(D|L,S,C)$ showing that D is conditionally independent of L and S given C, if the value of C is known, information regarding the variables L and S will not affect D. Given any sequence of variables on any network, it is assumed that (if the parent nodes are known) two nodes that are not directly linked, are conditionally independent (Nadkarni & Shenoy, 2004).

A further assumption is that generally, the conditional dependencies used in BNs work under the assumption of stationarity. This means that the moments (the quantitative descriptors of the data) regarding the distribution of a variable are uniform. Generally, the stationarity that is assumed in practice is second-order stationarity, which means that the samples used for a study are part of a consistent mean and variance for properties in the study area (Webster, 2000) However, in instances where this assumption does not hold true, possibly while using time series data, it is possible to use a non-stationary Dynamic Bayesian Network, which allows the network structure and conditional probabilities to change over time (Robinson & Hartemink, 2010).

There are also a number of practical constraints involved in determining the model structure, as BNs cannot account for cycles or feedbacks (Jensen, 2001), hence the graphs are described as acyclic. Furthermore, there should be no more than four 'layers'

to the model structure to avoid unnecessary propagation of uncertainty (Marcot et al., 2006). The size of the CPTs at each node is given by:

$$Size(CPT) = S\prod_{i=1}^{n} P_i \tag{3-8}$$

where S is the number of states and P_i is the number of states in the ith parent node (Chen & Pollino, 2012). Therefore, if a node has many parents, its CPT will become very large which makes populating the table difficult, due to increasing demands for data (from either empirical data or a multifarious process of expert knowledge elicitation). If the conditional probabilities at each node are derived from data, Cain (2001) suggests that at least 20 cases for every combination of variables are required to ensure the model is robust.

Both the qualitative (model structure) and quantitative (conditional probability tables) parts of the network are used to make probabilistic inferences, which is how evidence is propagated around the BN, that is to say, how evidence entered in nodes (usually in the form of data) comes to influence nodes of interest (what is being predicted) (Pearl, 1988). It is possible to make inferences about any variable within the Bayesian network. Before any evidence (data) is entered into the network, the CPTs give the prior probability of each variable. When evidence is entered, this changes the CPT at a node from a probability distribution to a definite state (in the case of hard evidence). For instance, if the Land cover in Figure 3-1 is known to be arable, the probability of it being arable becomes 1 and the probability of any other state becomes 0. Once evidence is entered (for either a single or many nodes), the network is updated to reflect what has been learned. Now the conditional probability tables of the nodes of interest (i.e. those being predicted) shows the posterior joint probability distribution.

3.1.2 Forming a Network

There are three key tasks in constructing a BN; 1) Identifying relevant explanatory variables for the system component (soil property) being considered, 2) defining the relationships between these variables and 3) representing these relationships via a set of conditional probabilities (Kuhnert & Hayes, 2009). For the first task, it is necessary to undertake a comprehensive review of the existing knowledge about the property of interest (Chen & Pollino, 2012) in order to identify variables known to affect the property. In the case of the spatial mapping of soil properties, however, there are practical constraints on the variables which can be used. In order for the BN output to be represented visually (mapped), it is usually necessary to represent nodes spatially in the form of GIS data layers (Johnson et al., 2012).

Once the variables of interest have been identified, the relationship between them must be determined. This requires the construction of a conceptual model which links the 'driver' variables with a wider suite of environmental variables and outlines key assumptions inherent in the model about the relationships between drivers and the property of interest. Described as the qualitative part of the network (Nadkarni & Shenoy, 2004), the most simplistic Bayesian model structure is known as a Naive Bayesian Network or a Bayesian Classifier (Duda & Hart, 1973). Here, the structure is very simple (Figure 3-5), as the naive network works under the assumption that all variables (L, S, C) are independent of each other given information about the root of the network (Bulk Density), which is the variable being predicted (Friedman et al., 1997). Note that despite all the variables (L, S, C) being used to predict D, the directional arrows run outwards from D. This is to allow the assumption of conditional independence.

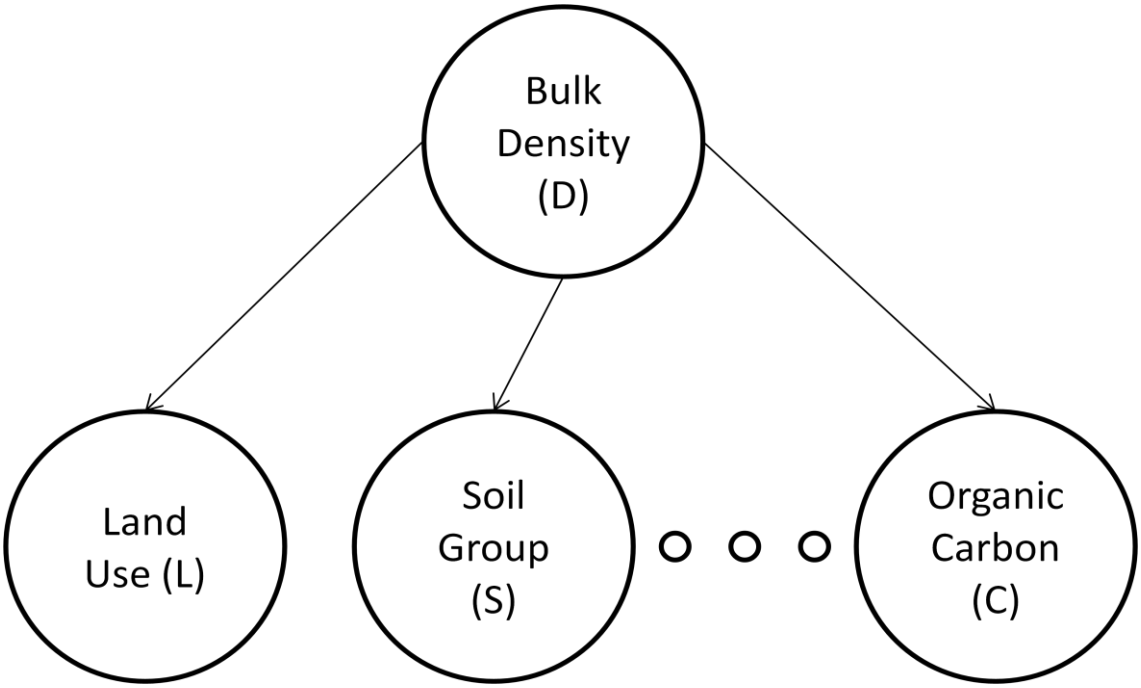


Figure 3-5: An example of a Naive Bayesian Network (adapted from Friedman et al., 1997)

A naive network is often used if there is little understanding of the system which is being modelled. For a more realistic representation of the system in question (in digital soil mapping, this will usually be the natural environment) it is desirable to model the interactions between variables in the network. This can be accomplished using a data-mining approach, frequently a Tree Augmented Naive Bayes (TAN) algorithm is used to derive the optimal structure of the BN (Friedman et al., 1997). The TAN algorithm modifies the naive network by identifying dependencies between predictor variables (L, S, C). The model structure is altered so that the predictor variables can have an additional parent node from one of the other predictor variables, based on the conditional mutual information contained in a training dataset (Jiang et al., 2005). While this approach has been shown to outperform naive BNs, the drawback is that it requires a large amount of data and predictions can be highly sensitive to changes in the model parameters. Furthermore, the complexity of environmental systems sometimes prohibits

these algorithms from producing adequate (feasible and efficient) model structure, such that superfluous nodes are included which can overcomplicate the network and reduce sensitivity to variations in relevant nodes (Chen & Pollino, 2012). In modelling, this is typically referred to as overfitting the data. BNs benefit from modelling parsimony, meaning it is preferable to exclude peripheral variables (those with little predictive power) to improve the ability of the model to predict independent data (Borsuk, 2008). The alternative is to construct the conceptual model using expert knowledge, where an expert or group of experts select the explanatory variables which are most likely to influence the predicted property and specify the relationships between them. At the very least, it is wise to incorporate an expert review of conceptual models built using a structured learning algorithm.

Once the structure is in place, the relationship between linked nodes can be defined by populating the CPTs via the application of the Bayes rule based on either empirical data, expert knowledge or a combination of the two. Before this stage can be completed, often some model parameterisation is required.

3.1.2.1 Discretization

Although it is possible to use continuous data for the variables in a BN, often these nodes are discretized into categorical data. This involves 'binning' continuous observations to create a series of discrete values. The motivation for doing this can be out of necessity, due to the algorithms used to calculate the conditional probability distributions, or because it is desirable as it will reduce the complexity of the (CPTs) within the network (Kuhnert & Hayes, 2009). Furthermore, it is easier for both experts and modellers to work with, understand and explain data which has been simplified in this way (Liu et al., 2002). In the discretization process, there are two main

considerations; into how many distinct groups the data should be divided and what the boundaries of these groups should be.

The two most straightforward methods of discretization are equal width and equal frequency binning. In equal width, the range of the continuous variable is divided into equally into a predetermined number of bins. In equal frequency, an equal number of observed values are place in each bin, subsequently determining boundaries. There are, however, drawbacks to these uncomplicated approaches as equal width discretization is particularly sensitive to outlying observation and in equal frequency, without post-discretization adjustment, it is possible that observations of equal value can be placed in different bins. There are also numerous statistical splitting and merging techniques which have been tested extensively on a range of datasets (Das & Vyas, 2010, Liu et al., 2002). These algorithms attempt to find natural breaks in the data, typically based on entropy reduction, and make the splits accordingly. The issue with adopting this approach for BNs where expert input or interpretation may be sought, is that it is recommended that discretized variables should have no more than five states (Marcot et al., 2006). This is certain to limit the benefits of using this method of discretization, hence generally, if data is sufficient, an equal frequency discretization approach will be adequate, especially as errors due to scaling or discretization will be more considerably less pronounced than errors in model structure (Druzdzel & van der Gaag, 2000).

3.1.2.2 Model Uncertainty and Evaluation

Frequently, BNs are not empirically validated, as they are often used to model scenarios where empirical data is scarce (Aguilera et al., 2011). In the absence of data to test the model, there are other methods to ensure the BN is as robust as possible. An important method of evaluating a BN is sensitivity analysis, used to determine which variables

within the network are most influential. Sensitivity is measured by the reduction in entropy or variance (depending on whether the node is discrete or continuous, respectively) of a target node when the model's parameters are varied systematically. In a BN, the values of the CPTs are varied and the effect on the probability distribution of the target node is recorded (Coupé & Van Der Gaag, 2002). This measure allows an expert to reassess both the model structure and CPTs if the findings are not what they expected (Chen & Pollino, 2012). Entropy reduction, as described by Marcot et al., (2001) is given by

$$I = \sum_q \sum_f P\,(q,f) \log \left[\frac{P(q,f)}{P(f)}\right] \tag{3-9}$$

Where I is the reduction in entropy, of target variable Q attributed to finding F. Here q is a state of target variable Q, f is a state of finding variable F and $\sum_q \sum_f$ are the sum of all the states q and f for the variables Q and F, respectively.

3.2 Modelling Soil Bulk Density

This study assesses the utility of BNs for predicting D_b at the landscape scale. In a broader context, this will go some way to establishing whether BNs can be used for a host of other digital soil mapping applications. The aim is to assess the extent to which BNs can be used in combination with readily available, landscape-scale data to produce physically interpretable models, which link soil D_b to easy-to-obtain environmental variables. The conditional probability distributions tested in the models are empirically derived, across a number of model structures in order to produce spatial predictions of topsoil D_b at the landscape-scale. Model inputs were selected as to be explicitly not reliant on point samples (with the exception of measured D_b data). One advantage of the approach taken is that the results can be directly compared to those obtained from

different statistical models for D_b built and analysed using the same dataset (Taalab et al., 2012). Finally, a map of predicted D_b values is produced, without interpolation, giving a uniform and quantifiable level of accuracy for the entire landscape.

3.2.1 Study Area and Data

The study was conducted in a 18150 km² region of the English Midlands, selected due to the relatively high density of pre-existing D_b sample data (Figure 3-6). The soils in the area are dominated by brown earths and surface water gleys, most of which have either a coarse or fine loamy texture, with some more clayey soils in the south of the region (McGrath & Loveland, 1992). A total of 342 D_b samples from the A Horizon were used in this study collected between1970 and 1987 during the 1:25000 and 1:50000 soil mapping of England and Wales. Models were built using 239 training samples and validated using the remaining 103 samples. The other covariates used in the model, which were sampled in ArcGIS 10.1 (ESRI, 2011), are detailed in Table 3-1. A detailed description of the study area and data used is given in section 2.2.1.

Table 3-1: Spatial explanatory covariates used in all BNs for the prediction D_b

Name	Description	Number of classes/ Range
AAR	Average annual rainfall derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005	548 – 1347 mm y^{-1}
Aspect	Aspect derived from a 10m DEM (Childs, 2004)	-1 – 360 (Discretisized into 5 classes)

AT0_Annual	Average accumulated temperature above 0°C derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005)	2564 - 3871 °C (Discretisized into 5 classes)
Curvature	Surface curvature derived from a 10m DEM (Childs, 2004)	-74.8 – 66.4 (Discretisized into 5 classes)
Elevation	Elevation above sea-level derived from a 10m DEM (Childs, 2004)	-2 - 558.9 m (Discretisized into 5 classes)
FCD_MED	Median number field capacity days derived from average monthly reports from the UK Meteorological Office on a 5 km x 5 km grid (Perry & Hollis, 2005)	107-290 days (Discretisized into 5 classes)
Great group	1:250,000 scale National Soil map of England and Wales (NATMAP; Hallett et al., 1996) classified into soil Great Groups (Avery, 1980)	5
Iwahashi	Iwahashi landform classification uses a terrain classification algorithm based on slope, surface texture and local convexity (Iwahashi & Pike, 2007) derived from a 10m DEM	8
Land cover	Land cover derived from the 1 km x 1 km Land Cover Map 2000 produced by the Centre for Ecology and Hydrology (CEH) (Fuller et al., 2002)	14
LEX	British Geological Survey (BGS) 1:625,000 scale map detailing the lexicon of named rock units	63
PM1	Soil parent material derived from a 1:250,000 scale Soil map of England and Wales (NATMAP; Hallett et al., 1996)	18
Pennock	Pennock landform classification uses a terrain classification algorithm based on slope, curvature and catchment size (Pennock et al., 1987) derived from a 10m DEM	7

PSMD	Potential soil moisture deficit related to the balance between rainfall and potential evapotranspiration (Jones and Thomasson, 1985) derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005)	50 - 261 mm (Discretisized into 5 classes)
PT	Potential evapotranspiration is the amount of evaporation which would occur if water was not limited (Hess, 2000) derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005)	480 – 708 mm y^{-1} (Discretisized into 5 classes)
RCS	Bedrock geology derived from 1:625,000 scale British Geological Survey rock classification scheme map, detailing bedrock lithology	27
Slope	Slope derived from a 10m DEM (Childs, 2004)	0 – 74.9 (Discretisized into 5 classes)
Soil Association	Soils grouped to the association level (Avery, 1973) derived from a 1:250,000 scale National Soil map of England and Wales (NATMAP; Hallett et al., 1996).	24
STI	Sediment transport index derived from a 10m DEM	-67.4 – 0 (Discretisized into 5 classes)
SWI	Saga Wetness Index, a terrain-derived index of soil moisture derived from a 10m DEM (Böhner et al., 2001)	9.8 – 19.7 (Discretisized into 5 classes)

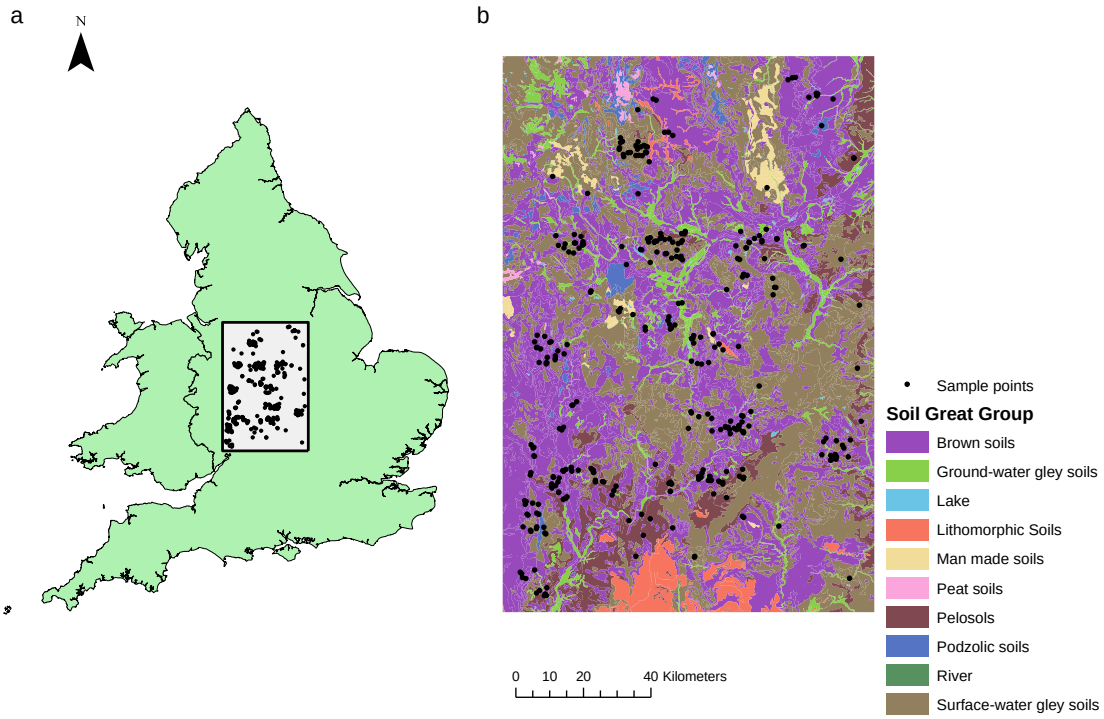


Figure 3-6: The Study Area. a) In relation to England & Wales b) Map of the Soil great groups within the study area (derived from NatMAP: Avery, 1980) and the sample locations (black points).

3.2.2 Model Development

Identifying the variables of interest was a relatively straightforward procedure for two reasons. First, a range of available, landscape-scale environmental variables have already been examined using linear (multiple linear regression) and non-linear (Random Forest and Artificial Neural Networks) modelling techniques within the study area, which has yielded information on the relative influence of different landscape variables on D_b (Taalab et al., 2012). Secondly, as the purpose of this study is to map D_b at a landscape scale, only environmental variables which can be represented at that scale (in the form of a GIS data layer) were considered as inputs (Johnson et al., 2012). When constructing the DAG for a network, it is important to explore a range of network

structures (Kuhnert & Hayes, 2009). For this reason, this study tests both naive and expert-derived structures, as well as a data-derived expert structure, using the Tree Augmented Naive Bayes (TAN) learning algorithm (Friedman et al., 1997). As well as the naive network, an 'optimised' naive network was developed using a stepwise classification procedure where each individual prediction variable was added to the model in turn and the predictive capabilities of the variable were assessed using the training data. To clarify, the predictive power of the network is tested using each individual predictor. The predictor variable which leads to the most accurate prediction is then added to the network structure (for example Land cover). The cycle then repeats for the remaining variables. A variable is rejected if it leads to an increase in the predictive error of the model, hence only predictors which lead to an increase in the predictive power of the model are included. The optimised naive network was built using the training data.

The nodes containing continuous variables were discretisized into 5 classes. After testing the two most commonly used discretization techniques (equal width and equal frequency) along with the 'minimum description length principle' (MDLP) algorithm (Fayyad & Irani, 1993), which, when tested on a range of datasets, has been shown to perform consistently well (Liu et al., 2002). Using a stepwise procedure, the study determined the percentage error in predicting D_b using a naive BN where the data had been discretisized into five classes using the MDLP, equal width and equal frequency techniques (Figure 3-7). The error results were generated using 'test with cases' feature of Netica (Norsys Software Corp, 2012), which determines the proportion of the training data which is correctly assigned to the correct class given the different class boundaries due to discretization. The purpose of this procedure was to assess the most

suitable method of discretizing data for use in the forthcoming models. Figure 3-7 shows that equal frequency discretization is the most suitable method for the task. This is consistent with the conclusion of Aitkenhead & Aalders (2009) that when some categories within a landscape are a lot more prevalent than others, a frequentist approach often gives a better representation.

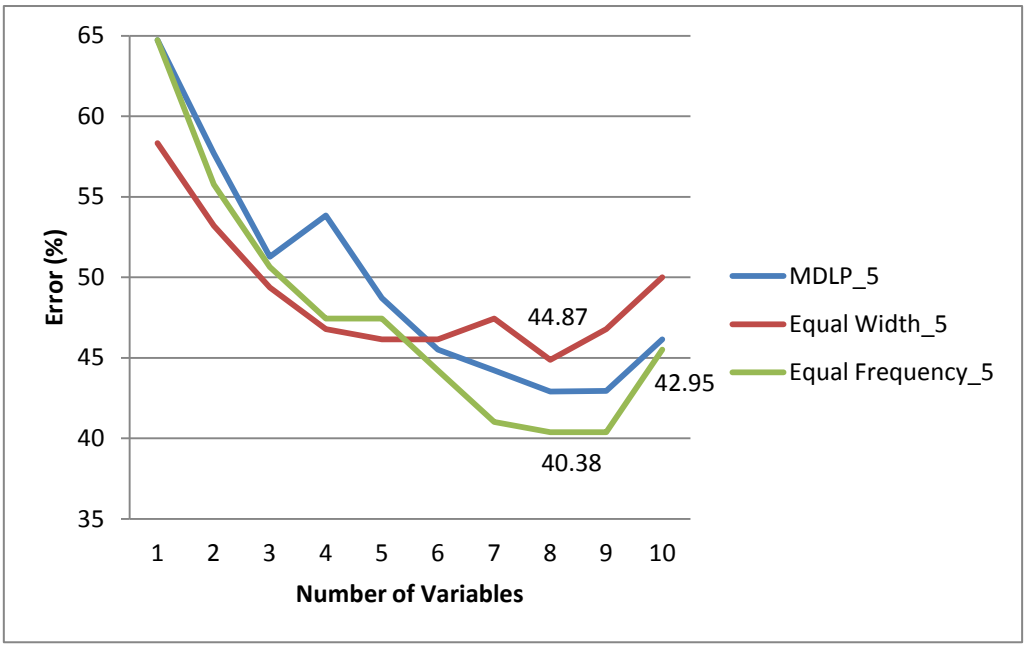


Figure 3-7: The relative error of the MDLP, equal width and equal frequency discretization techniques associated with predicting soil bulk density using a naive BN, where all continuous variables have been discretized into five classes.

The expert structure was determined collaboratively by two experienced soil scientists based at Cranfield University, Dr. R.J.A. Jones and Dr. J.A. Hannam. A number of model structures were discussed until a consensus (Figure 3-10) was agreed upon.

3.3 Results

3.3.1 Mapping Soil Bulk Density

The results (Table 3-2) show that the BN which was best able to describe the variation in topsoil D_b is the optimised naive network shown in Figure 3-8. This is a naive

network that uses an optimisation algorithm to identify and remove any variables which cause the prediction error to increase. The second-best performing model was the expert structure model (Figure 3-10). Although performance was similar to that of the naive network, notably fewer predictor variables were used. The most important variables determined by the 'Sensitivity to Findings' feature of Netica (Norsys Software Corp, 2012) (the five most important are ranked in order in Table 3-2), were Land cover, climatic factors, notably rainfall and soil association.

Table 3-2: Independently validated results of the each of the BNs

Network	R^2	RMSE	Variables
Naive	0.38	0.19	1. Land cover 2. Average annual rainfall 3. Potential Evapotranspiration 4. Median number of field capacity days 5. Average accumulated temperature above 0°C
TAN structure	0.34	0.19	1. Average annual rainfall 2.Bedrock geology 3. Profile curvature 4. Land cover 5. Slope
Naive Optimised	0.49	0.17	1. Land cover 2. Average annual rainfall 3. Median number of field capacity days 4. Soil association 5. Elevation
Expert Structure	0.39	0.18	1. Land cover 2. Soil Association 3. Saga wetness index (SWI) 4. Elevation 5. Average accumulated temperature above 0°C

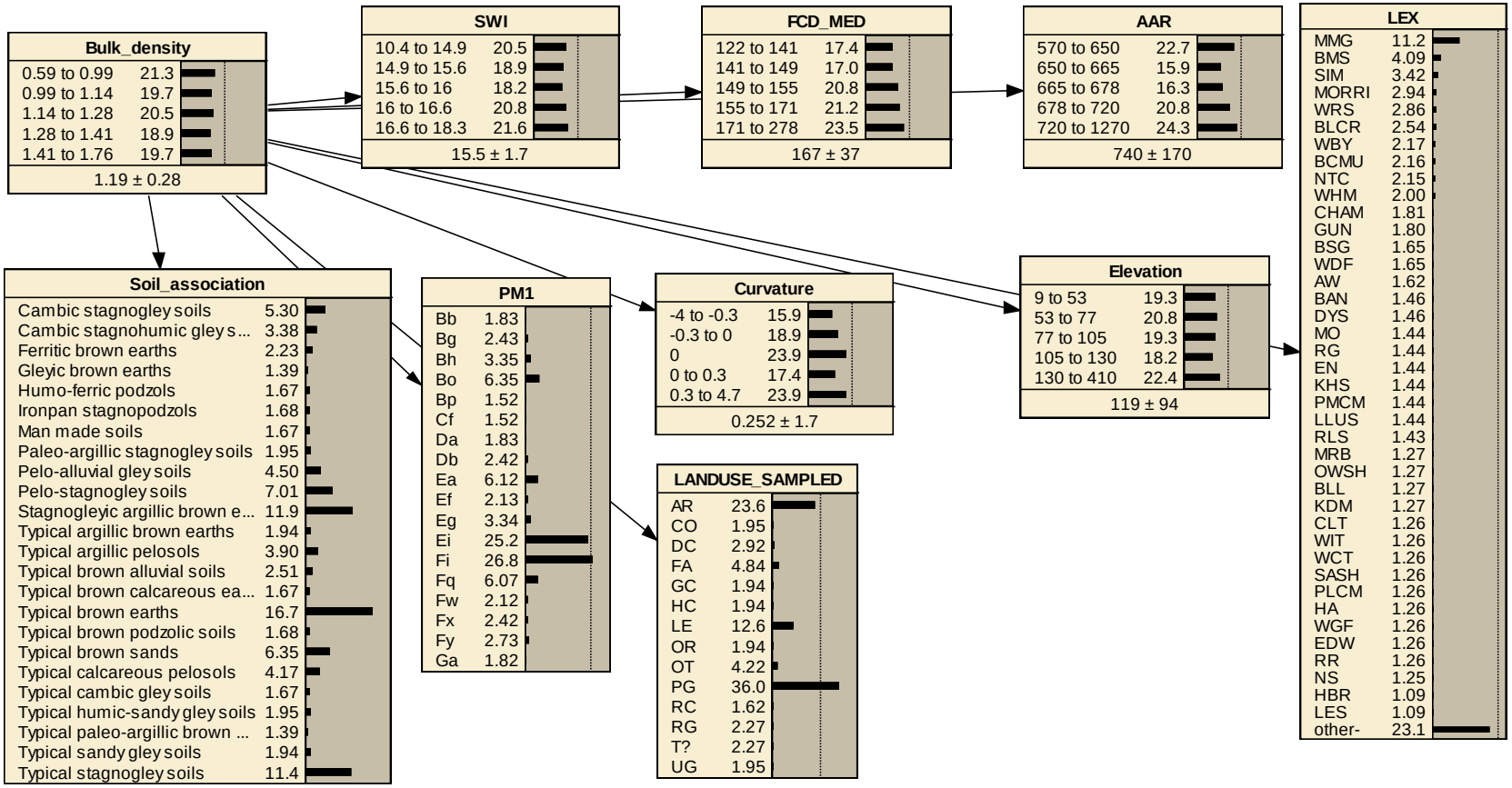
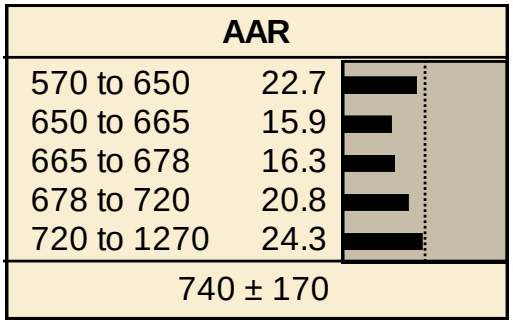


Figure 3-8: The optimised naive network. The variables included were determined using and optimisation algorithm selected only variables with significant predictive power based on the training data.



Bulk Density (g cm^{-3})	Average Annual Rainfall (mm yr^{-1})				
	570-650	650-665	665-678	678-720	720-1270
Very Low (Below 0.99)	7.14	21.43	8.93	8.93	53.57
Low (0.99-1.14)	15.38	13.46	19.23	26.92	25.00
Medium (1.14-1.28)	22.22	18.52	18.52	18.52	22.22
High (1.28-1.41)	24.00	16.00	22.00	28.00	10.00
Very High (Over 1.41)	46.15	9.62	13.46	23.08	7.69

Figure 3-9: An example of the conditional probability table (CPT) at a node, in this case Average Annual Rainfall (AAR).

Figure 3-8 and Figure 3-10 show the structures of the 'optimised naive' and 'expert-structured' networks. The bars displayed at each node are particular to the software used (Netica) and it is impossible to infer the relationships between variables by looking at the display alone. The bars represent the average probability of each of the states across all D_b classes. In order to understanding the relationship between (in this example) D_b and average annual rainfall (AAR), the CPT for the node can be examined (Figure 3-9). This shows the probabilistic relationship of the five D_b classes to the five AAR classes. The general trend reflected in the data is for Low D_b to occur in areas receiving the highest AAR (and vice versa).

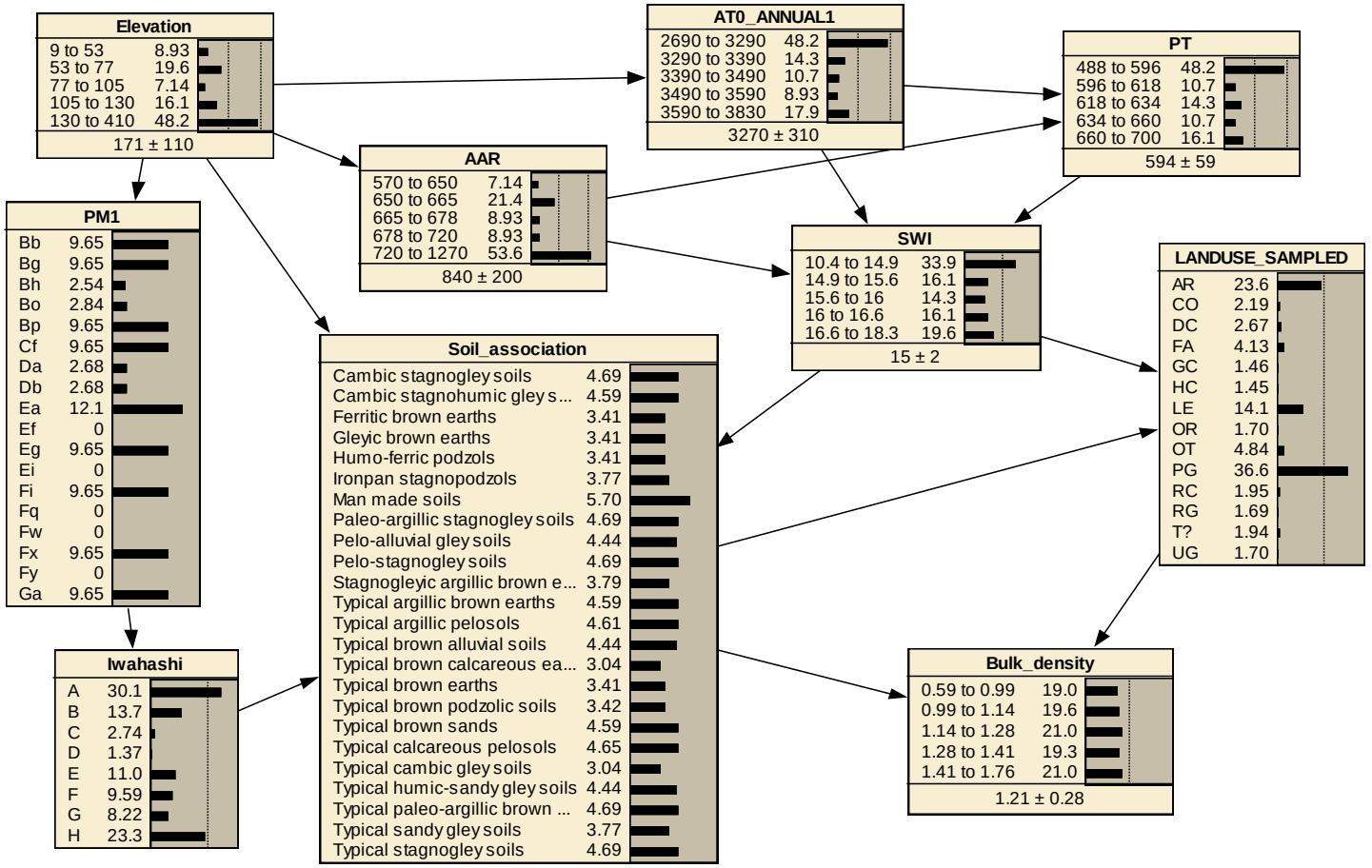


Figure 3-10: The expert-knowledge structured BN. The variables included and the links between variables were determined using expert knowledge.

Using the optimised BN, the best performing model, it was possible to create a continuous predicted surface of D_b (Figure 3-11). This model can account for nearly 50 percent of the variation in topsoil bulk density using the following landscape covariates;

- Land cover
- Average annual rainfall
- Median number of field capacity days
- Soil association
- Elevation
- Rock classification scheme
- Parent material
- Soil wetness index.

The covariates were listed in order of importance in making the prediction based on the Sensitivity to Findings analysis and measured by the reduction of entropy (3-7) (See Marcot et al., 2006).

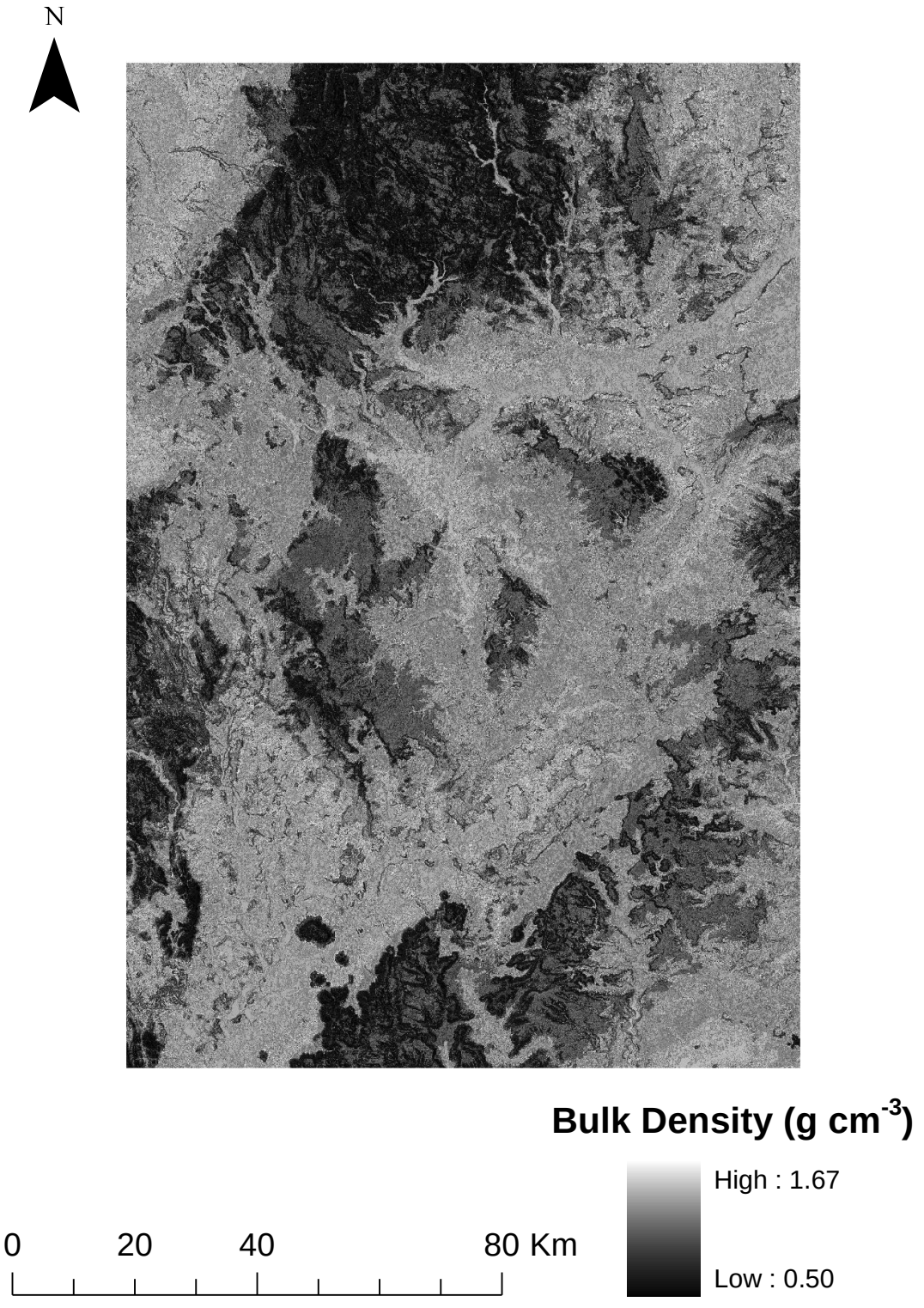


Figure 3-11: A continuous spatial prediction of D_b made using the optimised naive network.

3.4 Discussion

3.4.1 Model Performance

The performance of the optimised naive network is very similar to (albeit slightly lower than) the Artificial Neural Network and Random Forest black-box modelling techniques which have previously been used to predicted topsoil D_b with the same dataset (Taalab et al., 2012). While BNs did not improve predictive performance, they have the obvious advantage of offering some process-based insight (Correa et al., 2009). For example, the optimised naive network (Figure 3-8) shows that topsoil D_b is most likely to be very high (above 1.41 g cm^{-3}) in areas of low elevation and rainfall, on Brown Podzolic soils which overlie drift with siliceous stones. In conjunction with expert knowledge, this can either confirm or contradict the opinions of the expert(s). In this instance the results are plausible as areas of low rainfall would typically be associated with low organic matter and hence high D_b. Although Brown Podzolic soils would not necessarily be associated with the highest D_b values, Hallett et al. (1996) found that Podzolic soils can be associated with extremely high D_b values. If the BN contradicts what the expert believes, it can both prompt further investigation into the process, or possibly point to a lack of understanding. Alternatively, the problem may be with the model itself. If this is the case, it is easy to amend both the model structure and the probabilistic relationship between nodes. Identifying the source of predictive inaccuracies in a black-box model is much less straightforward. As they are based on process understanding, BNs can be used to answer specific questions using predictive reasoning. For example, what is the probability of 'high bulk density', given certain information, a capability that the black box models do not possess. Furthermore, BNs are also capable of diagnostic reasoning. For example, given an outcome, it is possible to predict favourable conditions likely to

lead to this outcome. These applications have already been applied to predict the locations of suitable habitat for endangered species (Smith et al., 2007) and more pertinently for the digital soil mapping community, to assess spatially the risk of peat erosion (Aalders et al., 2011).

That the best performing BN was a naive network is, at first, surprising as generally, the best BNs are those which combine an expert derived structure with a series of conditional probabilities calculated from measured data **(**Nadkarni & Shenoy, 2004). Many of the assumptions in a naive network can be unrealistic as they ignore correlations between predictor variables and hence do not represent real life accurately. Despite this, they have the advantage of avoiding superfluous dependencies of an over-complex network and are frequently found to be competent predictors (Friedman et al., 1997). The performance of the TAN BN is also of interest. Friedman et al. (1997) found them to be far superior to naive BNs, whereas this study found them to have the worst predictive performance. This can be attributed to the relative lack of data which, in our study, has led to overfitting of CPTs (the R^2 value of the predicted vs observed values of the training data was 0.86), meaning random error associated with these data was included in the model. This highlights the importance of testing models using independent data, to get an accurate estimate of a model's predictive power. Of the maps produced Figure 3-11 has similar spatial patterns to those produced by the RF and NN methods (Section 2.4.3) which reinforces the idea that BNs should be considered within the suite of data-mining models used for environmental mapping.

Although BNs explicitly model uncertainty, they are themselves subject to second order uncertainty. The uncertainty associated with BNs typically comes from inadequate datasets, bias or a lack of understanding within expert opinion and from imperfect

representation of real life by the model structure. There is, however, no way of distinguishing between the sources, which makes formalising this uncertainty itself, in the form of a probability distribution, uncertain. Hence, getting a genuine idea of model performance requires testing using independent data (Krueger et al., 2012). Often BNs are not subject to any validation (with the justification that the modelling approach is often applied specifically to situations where data are scarce). Aguilera et al. (2011) point out that, of the BNs which have been reported to solve regression or classification problems in environmental science between 1990-2010, fewer than thirty percent were tested using independent data. This is problematic for this type of modelling, because it will lead to BNs being compared unfavourably with other data mining techniques which are more routinely validated empirically. Jakeman et al. (2006) suggest that evaluation should go beyond the quantitative and include a subjective review of utility and transparency of the model.

3.5 Conclusions

Bayesian Networks provide a feasible alternative to black-box data mining techniques often applied to the modelling and mapping of soil properties. It is both their ease of interpretation and their ability to deal explicitly with uncertainty, which sets them apart. There are numerous approaches for the application of BNs to the prediction of soil properties many of which remain relatively unexploited (Aguilera et al., 2011). It is important to stress that that the cornerstone of good practice for the application of BNs is clarity throughout the modelling process (Chen & Pollino, 2012). A clear record of the choices and assumptions that underpin the model, in terms of parameterisation, model structure, elicitation and evaluation techniques is critical to ensure that the modelling approach remains credible. For digital soil mapping, BNs provide a logical

way of structuring knowledge, which can be used to disentangle complex processes, as well as 'filling in gaps' in empirical data. Many soil mapping applications are essentially attempts to formalise a soil surveyors' thought process, where the BN will perform expert-like reasoning. While this does not require expert opinion, as both the structure and CPTs can be 'learned' from data, all BNs will benefit from some form of expert evaluation to ensure that the relationships between variables are scientifically sound.

This study has demonstrated the effectiveness of BNs for quantitative prediction of a soil physical property (D_b) and qualitative prediction of soil class. In both cases the results were comparable to those obtained using black-box modelling techniques, with the benefit that the modelling process is easier to interpret. The study of soil D_b has shown that, where possible, model validation using independent data is invaluable. This is because expert judgment will always contain a measure of uncertainty reflecting both knowledge gaps and inherent natural variation which are difficult to separate. Current limitations in the ability of BNs to make highly accurate spatial predictions is offset by the clarity of the modelling approach (the process by which predictions are made) and the ability to model future scenarios (e.g. for different Land cover or climate regimes: Chen & Pollino, 2012).

4 The Application of Expert Knowledge in Bayesian Networks

This chapter investigates expert knowledge as a resource for digital soil mapping. To do this, three models of soil bulk density (D_b) were produced; i) a Random Forest model built and cross-validated using the limited data available (which served as the benchmark), ii) a naive Bayesian Network (BN), where the conditional probabilities defining the relationship between D_b and explanatory landscape variables were derived from expert knowledge rather than data and iii) an expert-derived BN which used a hierarchical structure. These three models were used to generate spatial predictions (maps) and populate indicative D_b values for soil taxonomic units for the 1:250,000 scale national soils map for Ireland. The expert knowledge derived maps were able to identify the same broad spatial trends in the variation in D_b as the Random Forest model. Furthermore, using both expert knowledge and data-mining approaches, it was possible associate D_b values with soil series providing a mean value and a 95% confidence interval, which could be compared with pre-existing reference values for each series. This demonstrates the potential for the use of expert knowledge as a proxy for empirical data, in situations where data is unavailable.

4.1 Introduction

There is a growing demand for high resolution digital information about soils, generally in the form of maps (McBratney et al., 2003). Mapping soil properties is particularly challenging in countries with limited quantitative data; many developing countries lack quantitative data but have an abundance of qualitative information in the form of soil surveys and classification studies (Hansen et al., 2009). This information is essentially a repository of expert knowledge, which has been used to in combination with limited

quantitative data to form rules for predictive soil mapping, generally with limited success (Stoorvogel et al., 2009). In this instance, a lack of empirical data restricts the ability of statistical models to generate robust rules for digital soil mapping applications. In DSM, a range of statistical approaches can be used to predict soil properties from readily available data. In practice, however, most studies rely on regression. Tree-based or geostatistical approaches are rare and knowledge-based modelling is implemented even less frequently (Grunwald, 2009). When expert knowledge has been used for soil mapping applications, it tends to be applied within a fuzzy logic framework (McBratney & Odeh, 1997; Zhu et al., 2001). However, probability theory can offer an alternative to this approach.

There is a growing acceptance and use of expert knowledge within environmental modelling. A key reason for this is the desire to make best use of existing knowledge in combination with available data to solve a host of environmental problems (Norton et al., 2012). Expert knowledge is defined as substantive information which is not extensively disseminated across the general population (Martin et al., 2012). It is garnered through a combination of training, technical skill and experience and generally an expert will be defined by the degree of their experience in relation to the topic of interest (Krueger et al., 2012). Those defined as experts, are expected to be able to recognise the most relevant attributes pertaining to a situation or problem (McBride & Burgman, 2012).

There has been a longstanding drive to formalise the inclusion of expert knowledge in the soil modelling process (Dale et al., 1989; Shi et al., 2009). Of the attempts to develop ‘expert-systems’ approaches to soil modelling, Bayesian methods have been identified as a potential method of structuring expert knowledge (Skidmore et al., 1996;

Bui et al., 1999; Corner et al., 2002; Farewell, 2010). Corner et al. (2002) used expert knowledge in a Bayesian framework to map soil attributes using 'Expector' (a custom-made Bayesian modelling software tool) and found it was possible to produce probability maps that a soil property falls within a predefined range (e.g. the surface clay content is between 5-10%) across a 200km² study area in Western Australia. On this basis, a final 'most probable' map was produced to map the spatial distribution of classes across the area. While direct assessment of the probability map is not possible using point samples, the performance of the 'most probable' map can be quantified. When validated using 200 sample points, 51% were found to be classified correctly which was an improvement on classes derived from a pre-existing soil map which classified 40% of the samples correctly. One caveat with these results is that the data used for validation was not independent as the same data was used to provide the initial probability estimates that were subsequently amended by the experts.

Bayesian networks (BNs) can be constructed using either data mining approaches (Heckerman, 1997), or knowledge-based approaches (Murray et al., 2012). This means it is possible to apply BNs to study areas where there is a shortage of empirical data. In a BN, expert knowledge can be applied to derive model structure (how nodes link to one another), the interactions between variables (the conditional probabilities at each node) or both.

In a DSM context, BNs are an attempt to formalise an expert's thought process behind decision making and hence to make the model perform expert-like reasoning (Krueger et al., 2012). Consequently, the model can be considered to be "process based", since experts will usually base their judgement about how different predictor variables

interact to drive the spatial variation of the property of interest on a conceptual perception of environmental processes. This is possible as the BN model structure is a more descriptive method of representing knowledge than many empirical approaches (Nadkarni & Shenoy, 2004). This means that a BN can be used as a way to structure and communicate an expert's hypotheses and will contain assumptions that are explicit. These assumptions can be seen in the model structure and within the conditional probability tables at each node. As these assumptions are based on opinion, they are inevitably subjective to some extent and, therefore, contain uncertainty. Usefully, this uncertainty is also captured in the probabilistic relationships between variables.

Another use for expert knowledge is for model evaluation. In DSM, soil maps produced using statistical models can be compared to those produced by expert knowledge which can be used to identify any discrepancies in the spatial distribution of a property of interest, for instance soil carbon stocks (Razakamanarivo et al., 2011). Moreover, experts can be used to moderate the model inputs to ensure that the variables used to generate predictions can be justified and are consistent with our current understanding of soil processes (Lark et al., 2007). For this reason, the information generated by expert knowledge needs to be both usable, meaning it has to be in a digital form and communicable, meaning that it must be able to be understood by whomever is the intended user(s) of the information. It should also be able to be validated and amended by other experts. In other words an expert system should be interpretable (Qi et al., 2006).

It should be noted that expert knowledge need not be provided by experts directly. It is possible to include expert knowledge in DSM applications on the basis of published

materials regarding soil-landscape interactions. This means that it is possible to apply an expert-derived mapping approach to areas which have sufficient legacy data (qualitative or quantitative) irrespective of the availability of current expertise. This can be particularly useful in data-poor environments, such as in much of the developing world (Hansen et al., 2009).

The objective of this chapter is to determine whether expert knowledge is an adequate substitute for empirical data for the spatial prediction of soil properties. To investigate this hypothesis, this study uses soil bulk density (D_b) as an exemplar soil property predicted using a Bayesian Network (BN) framework for expert knowledge. Although it is a key input to the calculation of many important soil properties such as carbon stock estimates, information on soil bulk-density is scarce and hence often predicted. Specifically, this chapter attempts to predict D_b at the landscape scale in the counties of Waterford, Kilkenny and South Tipperary in the Republic of Ireland, as a case study. This chapter also considers the degree to which 'representative' D_b values, obtained from soil pits representing soil taxonomic units (associations), can be predicted using these soil landscape models. This is to determine whether this particular approach can be used to populate 'representative' values of a soil taxonomy which is quantitatively incomplete.

4.2 Materials and Methods

4.2.1 Study Area

The study area comprises a 6162 km^2 region in Southern Ireland, made up of three counties: Waterford, Kilkenny and South Tipperary. The soils in the north of the study area are dominated by minimal grey-brown Podzolic soils while those in the south are

primarily acid brown earths. There is also a sizeable band of podzolic and peaty podzol soils across the central region (Figure 4-1).

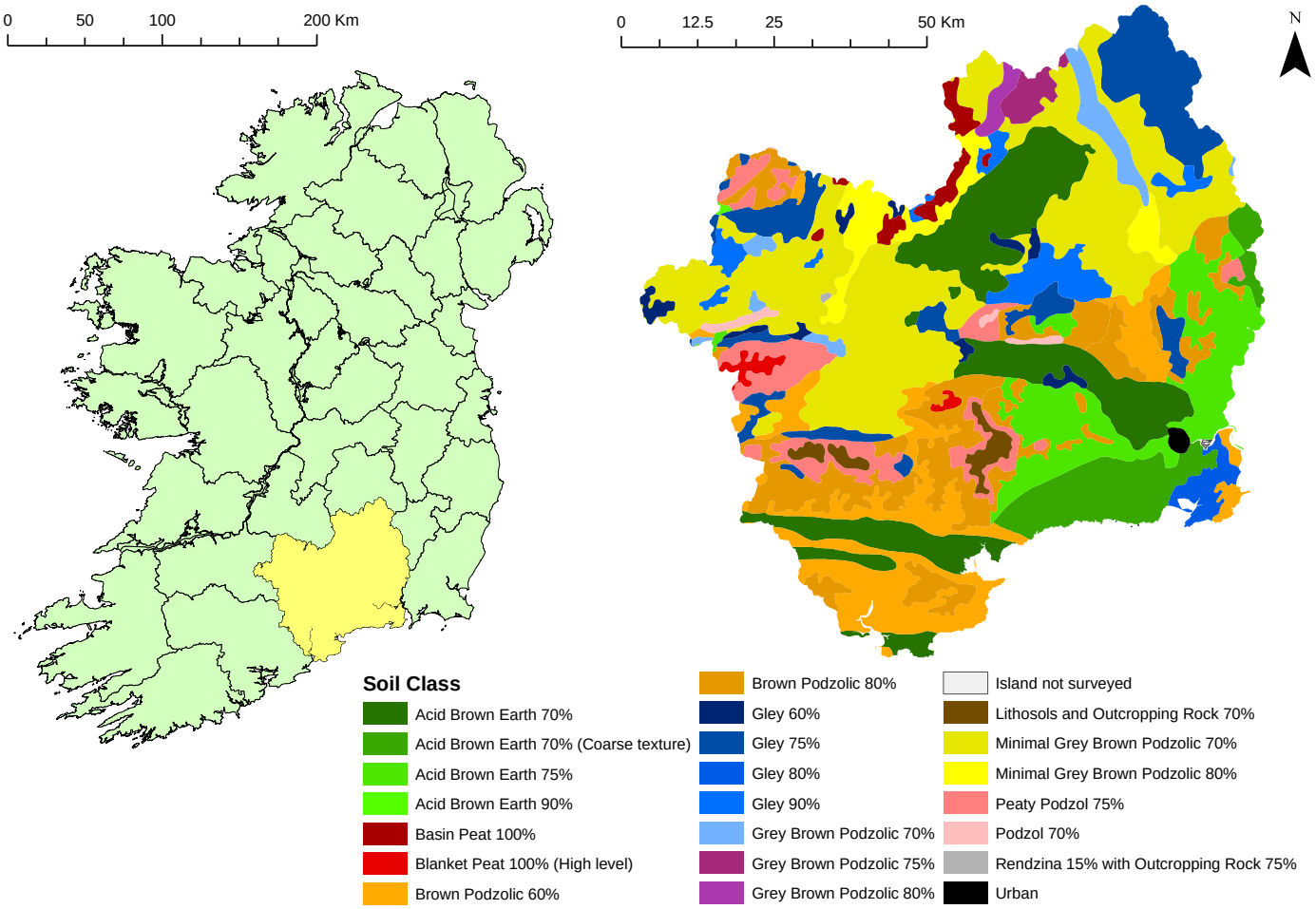


Figure 4-1: Study area. Showing the study area in relation to the rest of Ireland and the dominant soil groups within the study area

The land cover in the region is dominated by pastures and there are several distinct areas which are covered by peat bogs. The bedrock geology of included in the study area is complex and not dominated by a single class. The two most prevalent types of bedrock are limestone and sandstone, which generally tend to be found in the north and south of the region, respectively. There are also sizable areas of metamorphic and igneous geology. The elevation of the area ranges from sea-level to over 900 m above sea-level with a mean of 127 m, while the average annual rainfall ranges from a low of 842 mm to a high of 2,362 mm. The spatial distribution of these properties is shown in Figure 4-2. This suggests that the environmental covariates are not independent of one another. This is a noteworthy point as they will be treated as independent in a naive BN, an assumption this work acknowledges as incorrect. Despite this, the modelling approach is valid as a starting point, to help determine the relationship of each variable with D_b and assess the state of expert knowledge surrounding the subject. This approach can go on to inform a BN which attempts to represent the interactions between variables.

4.2.2 Random Forest Model

A Random Forest (RF) data mining model (Liaw & Wiener, 2002) was constructed to predict D_b using landscape scale explanatory variables (Taalab et al., 2012). The model was constructed in R using trained using 164 D_b samples and a suite of explanatory landscape variables (see section 4.2.3.2) and evaluated with 10-fold cross validation using the 'trainControl' function of the R package 'Caret' (Kuhn, 2008). A detailed explanation of the use of Random Forests is provided in section 2.2.3.2.

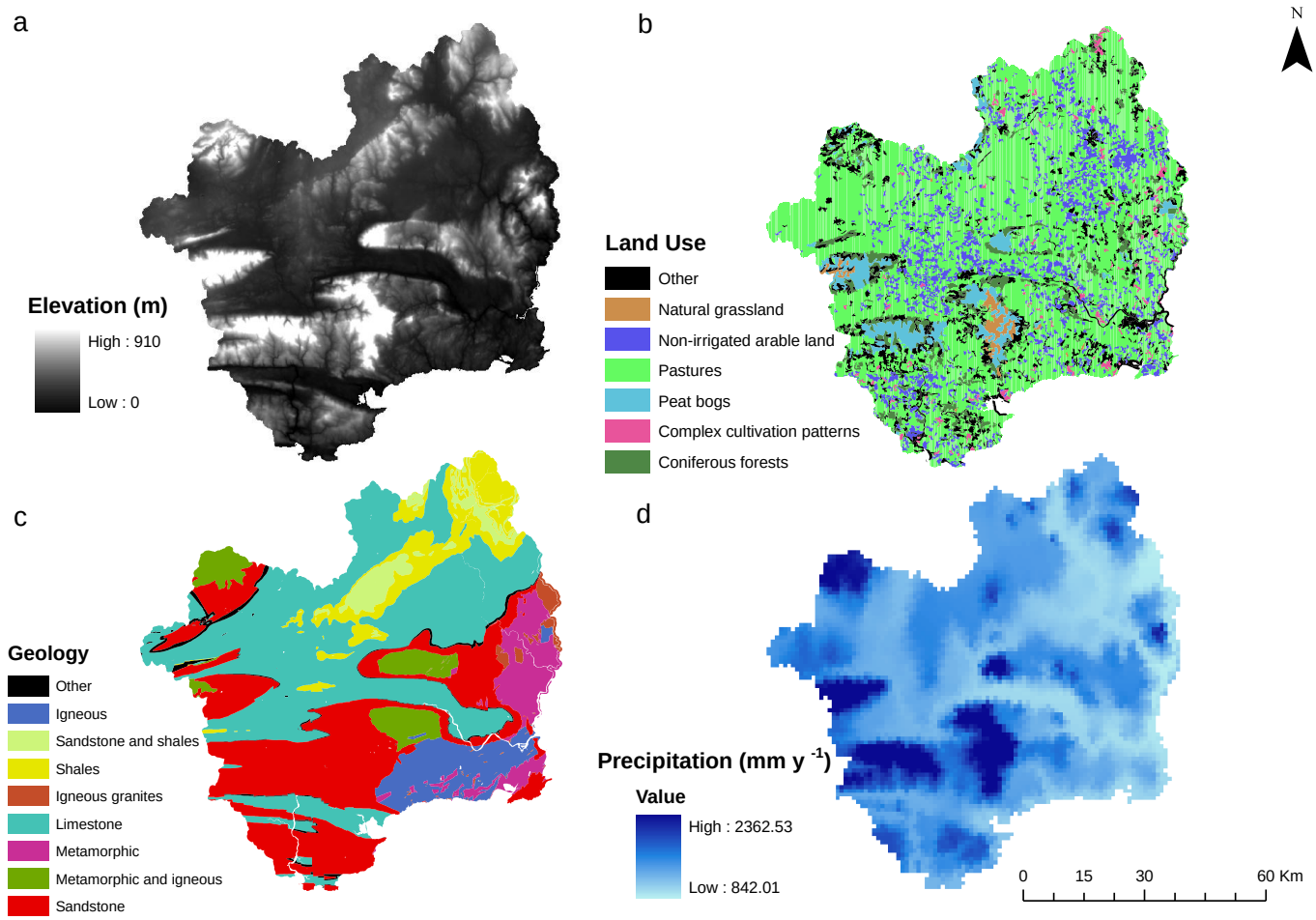


Figure 4-2: the spatial distribution of soil forming factors across the study area. a) elevation (m) b) Land cover c) bedrock geology d) average annual precipitation (mm y^{-1})

4.2.3 BN Model Development

4.2.3.1 Input Data

Soil taxonomic data were taken from the 1:575,000 scale General Soil Map of Ireland (Gardiner & Radford, 1980). This map shows 44 soil associations nationally, 22 of which are present in the study area (Figure 4-1). Subsoil data were taken from a 1:50,000 scale map produced by Teagasc and transformed into a new classification system containing 39 divisions nationally, during the ISIS Project. Eleven of the 39 divisions are present in the study area.

The topographic data in the study area were generated from a 20 m resolution digital elevation model (DEM) (Preston & Mills, 2002). As well as determining the elevation, this DEM was used to derive other topographic attributes. Slope was derived using the 'Spatial Analyst' tool in ArcGIS (McCoy & Johnston, 2002). The topographic propensity for soil wetness was represented using the 'SAGA wetness index' (Olaya & Conrad, 2009), which is based on the TOPMODEL Topographic Index (Beven & Kirkby, 1979) but uses a modified catchment area to create a more realistic representation of flow (Böhner & Selige, 2006).

Land cover data came from two sources: the CORINE land cover map is a 1:100,000 scale map Land cover map covering Europe, produced from an interpretation of satellite imagery sourced from the Landsat TM and SPOT HRV satellites (Büttner et al., 2002). There are 9 distinct CORINE Land cover classes in the study area (Figure 4-2b). The second source is a Teagasc-produced habitat map which is a 25 m resolution raster divided into eight classes in the study area. This map is expert-derived, based on land

cover satellite imagery taken from the 'Thematic Mapper' onboard the LANDSAT 5 satellite.

Data on the bedrock geology were taken from the Geological Survey of Ireland map at the scale of 1:100,000 (Naylor, 1978), which has been subsequently re-classified and harmonised during the ISIS project into new classification system for Ireland consisting of 28 divisions nationwide, 10 of which are present in the study area. The parent material was taken from the 1:575,000 scale General Soil Map of Ireland (Gardiner & Radford, 1980), 15 distinct classes are found in this study area.

The climatic indices were derived from long-term climatic records (1961-1990) supplied by Met Éireann and from the British Atmospheric Data Centre of the UK Meteorological Office. Data on rainfall and temperature were taken from 560 and 70 weather stations respectively across the country and averaged across a 1 km raster grid using polynomial regression (Goodale et al., 1998). As well as the average annual mean precipitation (Figure 4-2d) and temperature, annual mean Potential Evapotranspiration (PET) and Potential Soil Moisture Deficit (PSMD) are used as predictors. PET is a measure of how much water would be transferred from the surface to the atmosphere assuming no constraints on soil water supply (i.e. driven entirely by atmospheric conditions). This metric was calculated using the Penman-Monteith equation, as detailed in Hess (2000). PSMD was calculated using the monthly accumulated water balance deficit between precipitation and PET (French & Legg, 1979).

4.2.3.2 Training and Validation Data: Soil Bulk Density

The bulk density data used for this study comes from two of sources. Of a total of 164 D_b samples, 63 were collected from soilpits dug as part of the Irish Soil Survey

collected between 2009-2012 (Diamond & Sills, 2011). A further 101 D_b samples were collected over a 4 month period between September-December, 2010. In both cases, the D_b values were derived using the soil core method as described by Hodgson (1976). Briefly, soil cores, taken in triplicate were then dried at 110°C for 48 hours and D_b was determined using the methods described by Avery & Bascomb (1982). Triplicate measurements were treated as a single sample. Descriptive statistics on the D_b data are shown in Table 4-1.

Table 4-1: Measured soils data within the study area (n=164)

Variable	Mean	Maximum	Minimum	Standard deviation
Bulk density (g cm^{-3})	0.95	1.42	0.16	0.26

4.2.4 Expert Elicitation

As an expert-derived BN is aims to represent a probabilistic relationship between variables, it is reasonable to ask where the numbers come from. In short, they are made up by or in modelling parlance 'elicited' from expert(s). This is not problematic as it may seem as, frequently, expert knowledge matches measured data quite accurately (Spiegelhalter et al., 1990). However, for this to be the case, the elicitation process needs carefully planned. There are a number of guidelines regarding the elicitation process (Renooij, 2001; Garthwaite et al., 2005; Low Choy et al., 2009), which generally suggest implementing the following stages:

4.2.4.1 Problem definition and development of questions

The lack of spatial estimates of soil D_b is problematic in many DSM applications, most notably the prediction of carbon stocks (Grimm et al., 2008). As D_b is generally not measured, but predicted from soil textural properties, this typically confines prediction

to the point scale, however, landscape scale predictions are considered desirable for a number of applications (Chapter 2). In many regions, there is insufficient data to produce accurate statistical models describing the distribution of soil properties (Hansen et al., 2009), hence, mapping using expert knowledge rather than empirical data is of interest. As the final output of this expert derived BN is a map predicting soil D_b, A set of questions for experts was formulated to help create an expert-derived BN to map predicted soil D_b. Most of the questions concerned landscape-scale drivers of D_b, represented as GIS data layers. The variables used in the expert knowledge model are shown in Table 3-1.

4.2.4.2 Selection of experts

A major step in the elicitation process is the selection of the experts. Experts can be chosen on the basis of relevant experience, publication record, job and qualification. Also, if the knowledge is to be structured in the form of a BN then familiarity with probability theory is a bonus, although this is not essential and will be generally be uncommon. In reality, availability and (most importantly) a desire to participate are the two most vital traits (McBride & Burgman, 2012). In DSM, the scale of the study is another consideration. Soil mapping is generally quite a large scale endeavour, making soil scientists and soil surveyors a prime source of expert knowledge, however at a field or catchment scale, local farmers (for example) may have more detailed knowledge (Krueger et al., 2012).

Another consideration is whether to use a single expert or gather opinions from several. In terms of ease of implementation, using a single expert is preferable; however, associating an uncertainty to their prediction is less straightforward. For this reason, a using a number of experts is often preferable, generally, between 5-8 experts is optimal

(Clemen and Winkle, 1985). Within the group, diversity in terms of experience and training is desirable as shared training will often lead to similar beliefs. A further advantage to using multiple experts is that it guards against individual errors. Even an eminent expert may provide less accurate predictions that the collective knowledge of a group of less experienced experts (Clemen & Winkler, 1999). The expert knowledge was provided by a group of five soil scientists and soil surveyors ranging in experience from between 10 and over 30 years. All of the experts had substantial fieldwork experience within the study area and were familiar with soil D_b, its variability and the potential drivers for this variability. They were selected on the basis of experience, availability and willingness to participate. Although all the experts had some link to Teagasc (the Irish Agriculture and Food Development Authority), they had trained at a variety of institutions and had diverse professional backgrounds.

4.2.4.3 Question Format

Assessing probabilities is not a straightforward task, as experts can usually make relatively few probabilistic judgements about a variable (Garthwaite et al., 2005). To make the elicitation easier for the experts, the questions used to generate the conditional probabilities do not need to be posed in terms of probability, as it can be inferred from other information. Alternative approaches are especially useful when dealing with rare events, which would require very small numbers when expressed as a possibility. In this situation it is possible to other metrics such as odds, depending on the experts' familiarity with the concept. One popular method of assessing probabilities is to use a frequency format, where questions take the form 'given a set of conditions, from 100 samples how many would you expect to have situation X (Renooij, 2001). Kynn (2008) suggests using frequencies as a proxy for probabilities if possible, as they are typically

estimated with more accuracy. The reason for this is that is minimises certain biases such as over confidence, base rate neglect and the conjunction fallacy (Gigerenzer & Hoffrage, 1995) (5.5C.1). One drawback of this method is that it is not adept at expressing the probability of very rare events (Van der Gaag et al., 2002). These approaches are regarded as direct elicitation, alternatively, indirect approaches use words rather than numbers assess probability. In this chapter, probabilities were elicited directly, in terms of frequency using a questionnaire (Appendix 5.5C.2).

4.2.4.4 Training of Experts

During the elicitation, the expert(s) should be trained to think about problems in terms of probability. The format of elicitation for this study was two face to face workshop sessions, during which the experts were given a presentation detailing the heuristics and biases associated with the elicitation process. They were then given example questions to work through as a group in order to allow them to practice expressing their beliefs in terms of frequencies. Experts can practice elicitation by using questions sufficiently similar to the subject of interest (possibly using a different parameter where there is more data) and receiving feedback relating their answers to measured data (McBride & Burgman, 2012). In this instance, experts were give questions regarding soil organic carbon content using data from England and Wales. They were then introduced to the variables used in the modelling process and any questions regarding these variables, or the elicitation process in general, were addressed. The purpose of the training was to ensure that the experts were comfortable with the elicitation technique.

Table 4-2: Covariates used in the optimised BN

Covariate	Description	Source	Number of classes
COR	Land cover map	CORINE 2000 1:100,000 scale land cover map divided into 44 land cover classes countrywide, produced by interpretation of Landsat TM and SPOT HRV satellite imagery (Bossard et al., 2000).	7
elevation	Digital elevation model	Elevation model at a 20 m spatial resolution derived from interpolation of the contours of a 1:50000 Ordnance Survey map (Preston & Mills, 2002)	Discretized into 3 classes
GEO	Bedrock Geology	The Geological Survey of Ireland map at the scale of 1:100,000 re-classified and harmonised during the ISIS project into new classification system for Ireland consisting of 28 divisions nationwide	8
GSM	General Soil Map of Ireland 2nd Edition	1:575,000 scale, 44 soil associations countrywide presented at the Great Soil Group Level (Gardiner and Radford, 1980)	17
habitat1	Habitat class map	Teagasc-produced habitat map which is a 25 m resolution raster divided into 29 classes across the country (Fossitt, 2000)	8
Par_Mat	Parent material	The 1:575,000 scale General Soil Map of Ireland is combined into combined into 38 parent material classes across the country (Gardiner and Radford, 1980)	15
PET	Potential evapotranspiration	Potential evapotranspiration is the amount of evaporation which would occur if water was not limited (Hess, 2000) derived from average annual reports from the Irish Meteorological Office on 1km resolution grid	Discretized into 3 classes

Physio	Physiographic division	The 1:575,000 scale General Soil Map of Ireland is combined into combined into 9 broad physiographic divisions across the country (Gardiner and Radford, 1980)	6
precipitation	Annual mean precipitation	Annual mean precipitation derived from 560 weather stations extrapolated across 1 km resolution raster grid using polynomial regression	Discretized into 3 classes
PSMD	Potential soil moisture deficit	Potential soil moisture deficit related to the balance between rainfall and potential evapotranspiration (Jones and Thomasson, 1985) derived from average monthly reports from the UK Meteorological Office on a 1 km raster grid.	Discretized into 3 classes
SBS	Subsoil map	Subsoil map 1:50,000 scale map produced by Teagasc and transformed into new classification system during the ISIS project into 39 divisions	11
slope	Slope angle	The angle of inclination of the topographic surface derived from the DEM using the Spatial Analyst tool in ArcGIS (McCoy & Johnston, 2002)	Discretized into 3 classes
temperature	Annual mean temperature	Annual mean temperature derived from 70 weather stations extrapolated across 1 km resolution raster grid using polynomial regression	Discretized into 3 classes
wetness	Saga Wetness index	A terrain-derived index of soil moisture derived from the 20 m DEM (Böhner et al., 2001) using ArcGIS.	Discretized into 3 classes

4.2.4.5 Performing the elicitation

The elicitation process was performed in several stages, following guidelines based on the Delphi approach (Delbecq et al., 1975) proposed by Burgman et al. (2011a). The first was to allow the experts to discuss each variable in turn and assess how each might affect soil bulk density. The purpose of this was to identify any disagreements in the group and to identify whether these had arisen due to a linguistic misunderstanding that could be resolved immediately. A facilitator was present throughout the process, whose role it was to clarify these uncertainties and ensure that, where possible, the discussion was not dominated by a single expert. The experts were then asked to provide individual answers to the questionnaire detailing the frequency distribution of D_b for (Appendix 5.5C.2). These individual answers were then collated and presented back to the group. The experts were subsequently allowed to revise their answers or not, in light of group opinion. Individual responses were then aggregated using a mathematical 'opinion pool' approach, in this instance using the mean of the responses. Using the mean is a straightforward, robust technique, which performs as well as more complex aggregation techniques (Clemen, 1989). This elicitation was used to populate the Conditional Probability Tables (CPTs) of the naive BN (Appendices 5.5C.4 and 5.5C.5).

The first step in creating the hierarchical model was to create a conceptual model, representing the cause-effect relationships of the environmental variables used in the naive network. This process was carried out as a facilitated group discussion with the variables and links constructed using Netica software (Norsys Software Corp, 2012). The model went through several iterations until all experts agreed on the structure (Figure 4-3). On examination of the conceptual model, it was agreed that the conditional probability tables were too large and complex to be populated using expert knowledge.

Instead a more simplistic hierarchical model was proposed (Figure 4-4). The structure of this model is partitioned into 'layers' where landscape variables (such as soil class and parent material) are feed into a single 'soil' node which then goes on to influence D_b predictions directly. Structuring a BN in this way provides a more realistic representation of the processes acting on D_b while simultaneously keeping the CPTs small enough to be feasibly populated using expert knowledge (Murray et al., 2012). Constructing the hierarchical model involved identifying a much smaller number of variables on which to base predictions. It was decided that nine influential variables would be selected to represent three distinct factors affecting D_b: soil, Land cover and climate (Figure 4-4). The 'Soil' node consisted of soil group, subsoil group and parent material. The 'Land cover' node consisted of CORINE landscape classification, Habitat and landscape physiographic division (Gardiner and Radford, 1980). The 'Climate' node was derived from PSMD, PET and soil wetness index. The variables used in both the Soil and Land cover nodes were then categorised into four classes relating to their influence over D_b, ranging from 'high' to 'very low' based on a group consensus (Appendix 5.5C.3). The variables which formed the 'Climate' node were split into three classes. The reason for this difference is that both Soil and Land cover had an additional 'very low' class to account for the presence of peat soils. Once the network had been built, the CPTs were populated using a group consensus, rather than a mathematical average (Appendix 5.5C.6).

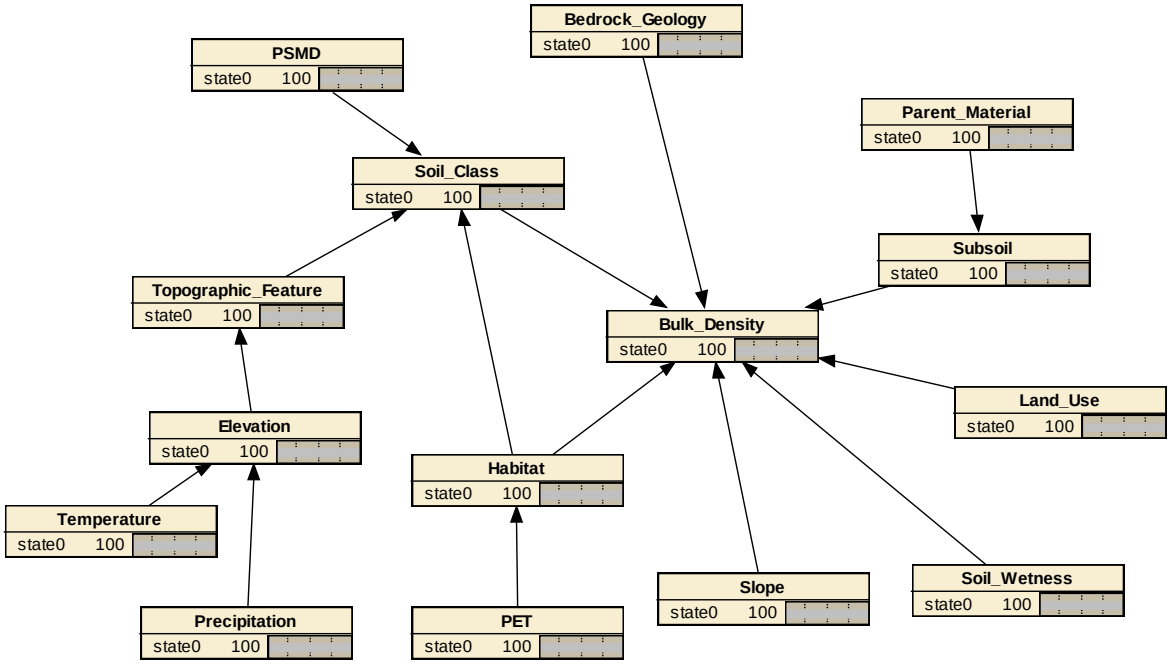


Figure 4-3: Conceptual model derived by the experts representing the cause-effect relationships between variables.

4.3 Results

Table 4-3: The results of the naive and hierarchical BN models. The sensitivity analysis ranks variables by how much changes at these nodes affects the D_b prediction, measured using reduction in entropy.

Model	R^2	RMSE	Sensitivity Analysis
Naive network	0.2587	0.2314	1. SBS 2.GSM 3.COR 4. Parent Material 5.Habitat
Hierarchical expert structure	0.4161	0.2268	1. Land cover 2. Soil 3. GSM 4. CORINE 5. Physio

Previous work has suggested that, with sufficient data, it is possible to predict around 55 percent of the variation in D_b using empirical data mining approaches for landscape scale prediction of D_b (Taalab et al., 2012). When a Random Forest data mining model was applied to the study area, 10-fold cross validation showed that the model performed

considerably below this level (R^2 = 0.39 with a RMSE of 0.1860). Although it is not a direct comparison, the hierarchical BN was able to explain just over 40 percent of topsoil D_b variation based on the validation results (Table 4-3). To clarify, both the BN models were validated using the same 164 measured D_b samples which were used to train the Random Forest model. For the BN models, this is independent validation as the data was not used to derive the conditional probabilities or model structure. For the Random Forest model, the results are produced using cross-validation as the same data is used to train and validate the model. This is the reason that the results of the two modelling approaches are not subject to direct comparison. In addition, the spatial patterns of D_b generated by both the naive and expert structured BN models are very similar to the map of D_b produced by the random forest model (Figure 4-7).

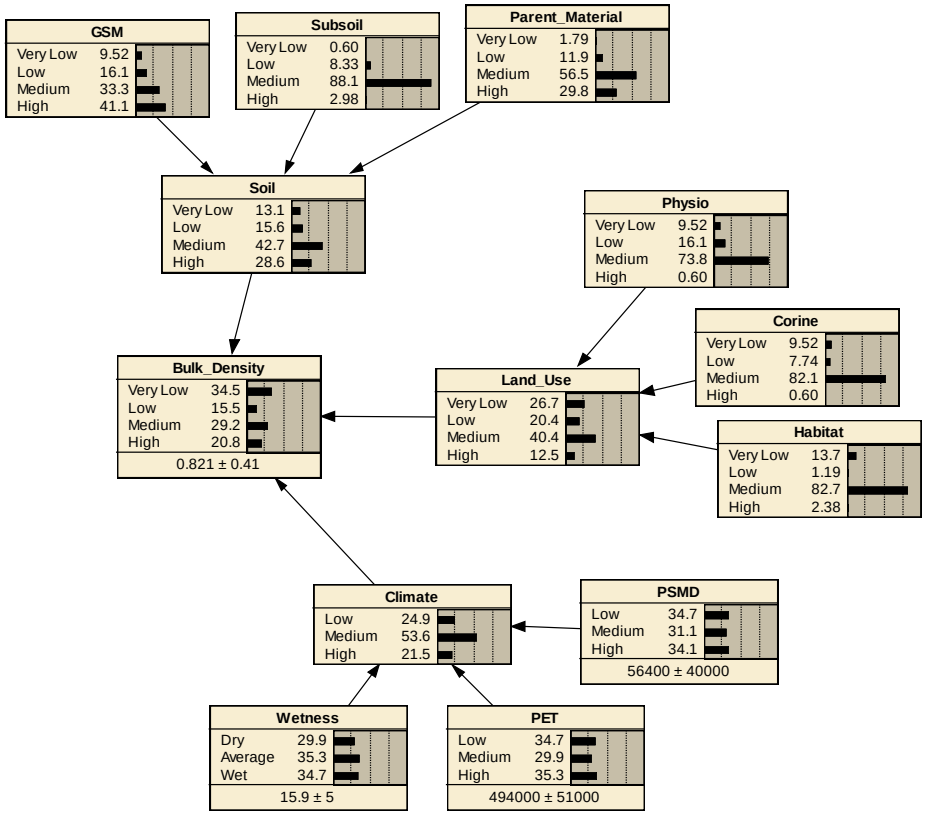


Figure 4-4: Hierarchical expert structured BN

Rather than comparing the predictive accuracy of an expert derived BN to a data mining approach, it is more accurate to compare it to the situation where no data is available, or rather data (in the form of expert knowledge) is not available in a coherent structure. The similarities in the distribution of D_b between the data mining and two BN approaches shows that at the very least expert knowledge can be used to identify large scale spatial trends in soil properties. This can be extremely useful for DSM applications, for instance in the design of soil sampling regimes. This suggests that a BN provides a useful framework for organising expert opinion into a useable resource. Generally, the best BNs are those which combine an expert derived structure with a series of conditional probabilities calculated from measured data (Nadkarni & Shenoy, 2004), hence, the CPTs can be augmented when and if new data becomes available.

While the Hierarchical model outperformed the Naive network, it did not explain the majority of variation in D_b. Nevertheless, expert knowledge still has value in identifying the 'big picture' key relationships between variables (Garthwaite et al., 2005). When tested against independent measured D_b data, the Naive network was found to generally provide very conservative estimates of D_b (most prediction were close to a mean value of around 1.1 g cm^{-3}) and was especially poor at predicting soils with low D_b (Figure 4-5). One reason for this is that when the organic carbon content of a soil is particularly (for instance in a peat or peaty podzol) the D_b will be low irrespective of the conditions of the other (potential explanatory) variables, such as Land cover or habitat. This was not reflected in the Naive network. It may be wise, therefore, to develop separate models for mineral and organic soils. This issue could also be resolved, in principle, by changing the structure of the network to reflect the interactions between variables more realistically. This is the purpose of the conceptual model (Figure 4-4). That said,

creating a BN directly from the experts' conceptual model would not have been feasible, as it would have required hundreds of conditional probabilities to be elicited for the bulk density node alone. Experts often tend to build conceptual models that reflect the complexity of the natural environment. It is the role of the model builder to identify the critical relationships between variables and to neglect those relationships which are likely to have less effect on the target variable, thereby keeping the model as simple as possible (Chen & Pollino, 2012).

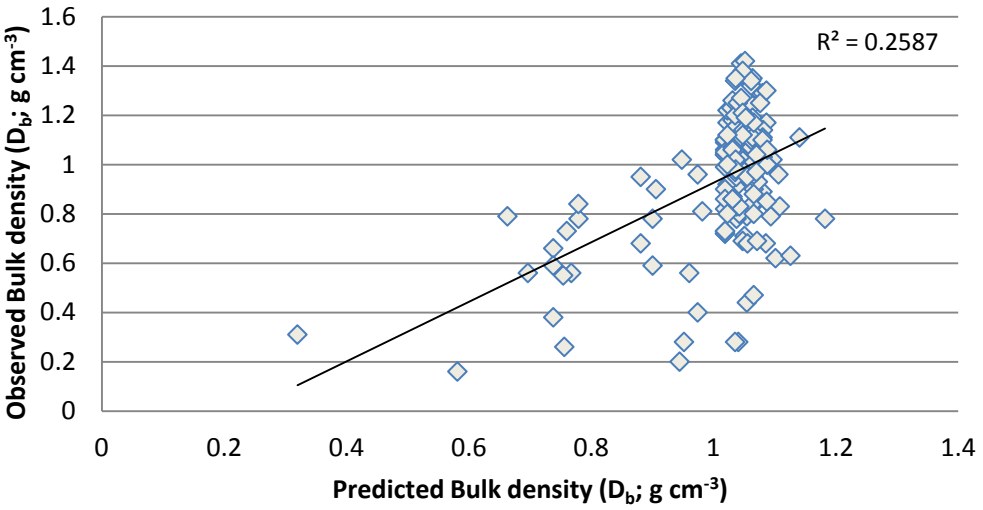


Figure 4-5: Predicted vs observed D_b values for the naive network results.

This prompted the construction of the hierarchical model (Figure 4-4) which made specifying the condition probability tables more straightforward. This model reduced the number of variables used for prediction and the number of parameters. This involved the reclassification of categorical variables (such as Land cover and soil group) into four D_b categories ranging from high to very low (Appendix 5.5C.3). The 'very low' category was established to allow the model to better predict the D_b of soils with high carbon content such as peat. Judging from the spread of residuals in the

hierarchical model, where the lower observed D_b values have generally low residuals (Figure 4-6). Thus, the adoption of a parsimonious BN model structure was relatively successful, statistically.

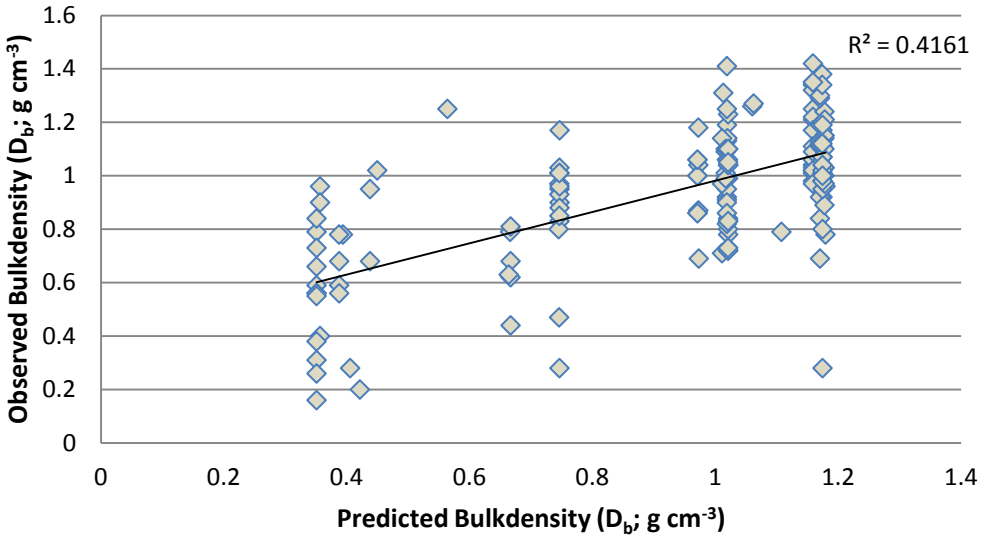


Figure 4-6: Predicted vs observed D_b values for the hierarchical expert structured networks.

In BNs, the potential reduction in predictive power caused by excluding peripheral variables, can be accounted for by amending the probability distributions at the other nodes to reflect the greater uncertainly associated with predictions (Borsuk, 2008). This simplified approach used in the Hierarchal model has been successfully employed in habitat suitability studies (e.g. Smith et al., 2007; Murray et al., 2012). While this model provided a better fit to the validation data, it could only account for just over 40 percent of the variation in D_b and had some relatively large residuals, especially for the medium D_b values. Figure 4-6 shows the predicted vs observed D_b predictions made by the Hierarchical model. The challenge of this approach is the amalgamation of several categorical classes into single D_b classes without significantly reducing the predictive

capability of the model. Large residuals (especially for medium D_b values; 0.8-1.1 g cm^{-3}) suggest that despite having mode predictive plover than the Naive network, it may be an overly-simplistic model to describe the majority of landscape scale D_b variation.

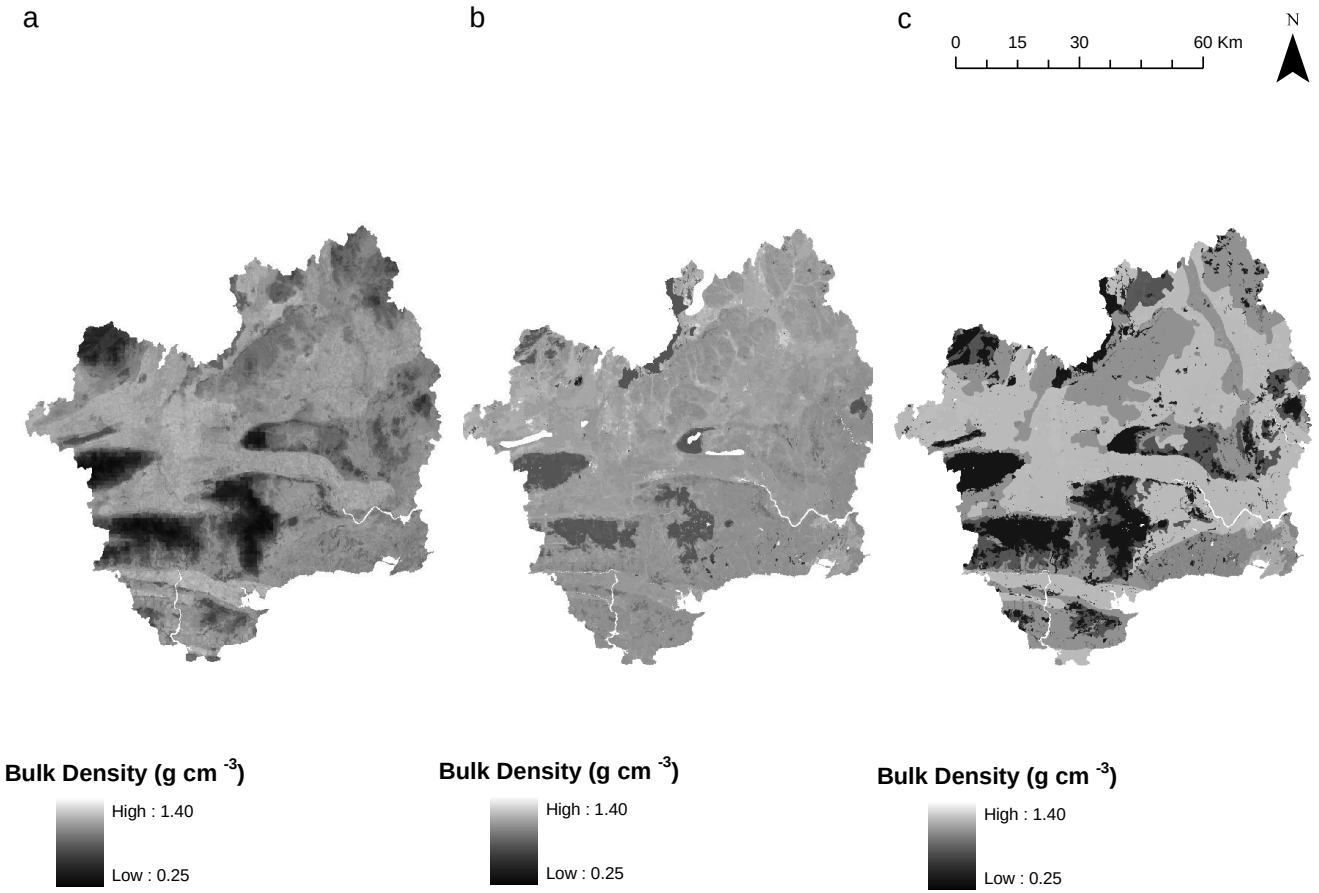


Figure 4-7: The spatial predictions of bulk density. a) Map produced using Random Forest b) Map produced using the naive network map c) Map produced using the hierarchical expert structured BN

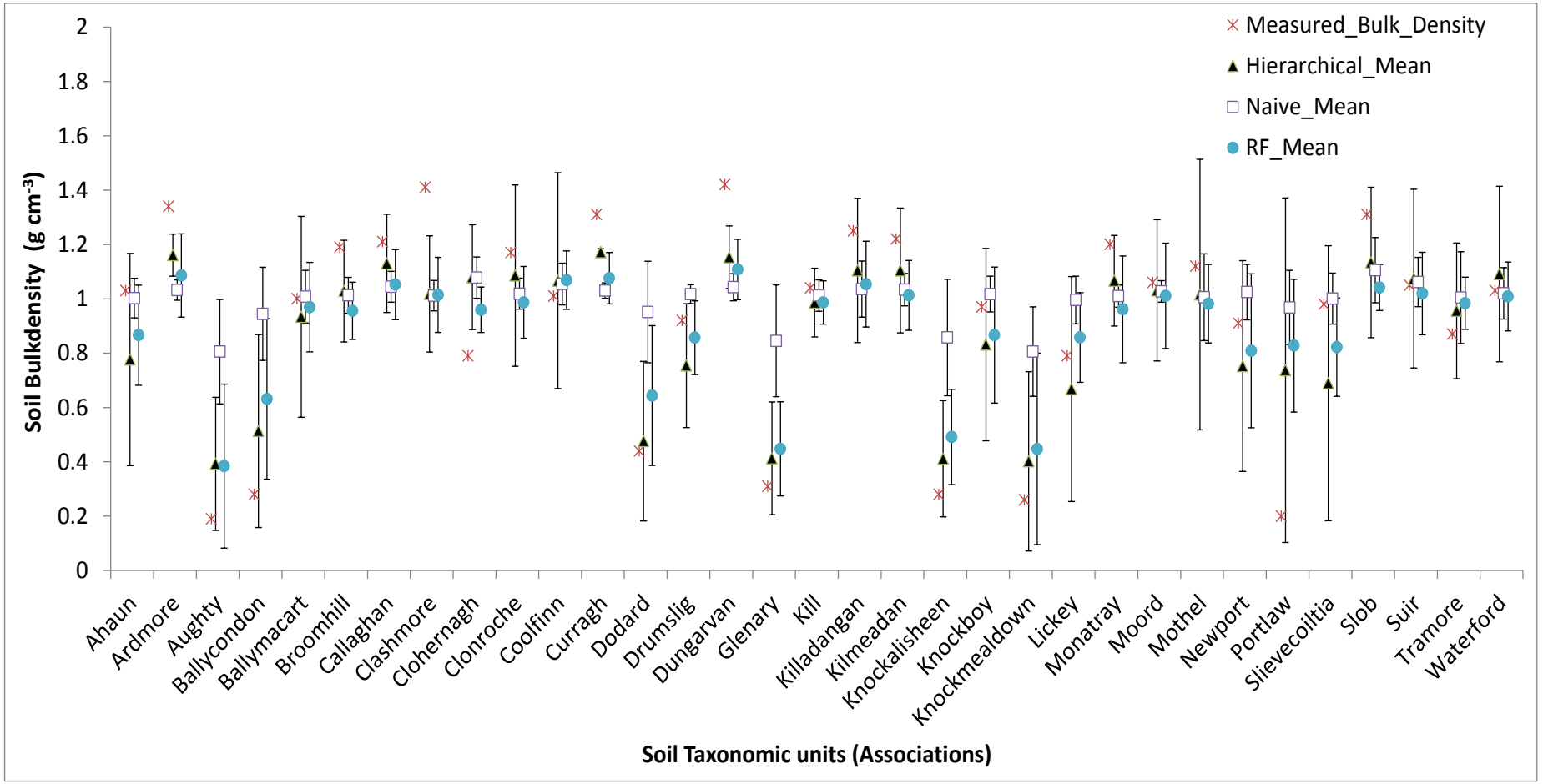


Figure 4-8: Soil bulk density predictions by soil associations

4.4 Discussion

4.4.1 Model Performance

One reason for the differences between BN model performance is that the expert knowledge was aggregated mathematically in the naive network and by consensus in the hierarchical model. This change was deliberate and was a direct consequence of the fact that the majority of D_b predictions made by the Naive network, fell in a very narrow range (Figure 4-5). This suggested that the range of probability distributions was too wide, and that the experts were being too cautious in their estimates. By implementing a group consensus, the experts were more inclined to include zero values rather than low probabilities, which has the benefit of reducing the complexity of the network (Jensen & Andersen, 1990). While this helps identify broad trends, it means some of the subtlety of the model is lost, which can result in clusters of predictions rather than a continuum (Figure 4-6). Both this, and the amalgamation of classes in the hierarchical structure results in maps with a more 'blocky' appearance (Figure 4-7c) and shifts the focus of map production to identifying broad spatial trends as opposed to more incremental change. Whether or not this is an appropriate approach depends on the soil property of interest and the purpose of the map being generated.

4.4.2 Elicitation Technique

On examination of the naive and hierarchical models, the CPTs of each are distinctly different, reflecting the mathematical averaging and group consensus elicitation techniques. In addition to considering how best to combine the opinion of a number of experts, it is worth considering the merits of the elicitation technique used. Estimating probabilities using the frequency format has the benefit of being intuitive for the expert

and is generally quite accurate (Kynn, 2008). Despite this, in light of both the BN models performance, it is worth considering whether alternative approaches may have provided more accurate results. An alternative approach used in a host of environmental models is first to elicit descriptive statistics comprising an upper and lower confidence interval and a 'best guess'. A statistical distribution can then be fitted to these bounds (O'Hagan & Oakley, 2004; O'Hagan, 2012). While this eliminates the need to elicit a full probability distribution and may be suited to other soil properties, it would be difficult to implement for soil D_b. In part, this is due to the fact that D_b varies over a fairly narrow range of values, meaning that predictions would need to be made for relatively narrow increments for which experts would feel uncomfortable to give precise numerical estimates (Renooij, 2001). Furthermore, D_b was a property of the soil that the experts were not used to quantifying in terms of g cm^{-3}, they had a sense of how it varied between soil types and Land covers, but were not comfortable giving numerical estimates. Another approach which may be worth considering would be to employ a graphical aid to help with classification. It is possible to use a GIS interface to help experts visualise the effect of their predictions across the study area. One potential drawback of this is that the experts may focus on the information they are most familiar with (e.g. soil maps) rather than consult the other available layers (e.g. topographic or climatic variables) (Yamada et al., 2003).

4.4.3 Experts

Soil systems are complex and non-linear with processes that vary greatly over different temporal and spatial scales. This means that it is likely that experts will sometimes be required to apply their knowledge beyond the limits of their experience (McBride & Burgman, 2012). This was the case with D_b, as prediction of this property is usually

based on soil textural properties and organic carbon content. To put D_b in the context of potential landscape-scale explanatory variables and to predict variation at a landscape scale required some lateral thinking by the experts. This is a balancing act between adaptive expertise (the ability of experts to apply their knowledge to unfamiliar scenarios) and extending an expert opinion beyond the limits of their expertise (Burgman et al., 2011a). As soil D_b is part of the soil survey handbook (Clarke, 1940) and is considered a fundamental part of soil structure it can be considered to be within a soil scientist or surveyor's knowledge domain. Predicting D_b, therefore, required the experts to apply adaptive knowledge which can, in fact, be inhibited by significant domain knowledge and experience because ideas about contributing processes become deeply ingrained, hindering the ability to relate the property of interest to an unusual context (McBride & Burgman, 2012). The fact that all the experts involved in this study had at least 10 years' experience may not have improved predictions in comparison to their less experienced colleagues. It can be the case that those with less experience are more able to adapt their thinking and, hence, better-predict new scenarios (Chi, 2006).

Although training and feedback are needed to improve expert predictions (McBride & Burgman, 2012), providing feedback may not lead to instantaneous improvements in expert predictions. Rather, feedback (such as 'calibrating' an expert's answers to the available data) tends to be more effective if given over time, allowing the experts to put their updated knowledge into practice (Burgman et al., 2011b). This can be problematic if there are time constraints associated with the investigation. In this study substantive feedback was not given for a number of reasons. Firstly, one of the aims of this study was to assess how useful expert knowledge was in the absence of quantitative data.

Secondly, with a limited dataset it is questionable whether it is valid to amend expert opinion to fit the empirical data (Kadane & Wolfson 1998). This could be construed as equivalent to overfitting a statistical model. The prediction of D_b is more likely to be subject to cognitive bias, in particular underestimation of uncertainty (especially in the hierarchical model) than motivational bias (Group thinking, wishful thinking) (Appendix 5.5C.1). In the study presented here, the creation of a map as an end product avoided some of the bias associated with the "stakeholder model", which can occur when the BN informs decision making. In the latter case, it is necessary to account for the motivations of the different groups which the decision will affect, which is something to be aware of in future work (Krueger et al., 2012).

4.4.4 Variables

An issue raised by the experts during the training phase of the exercise was the limitations placed on expert knowledge by limiting predictor variables to those which were available as GIS data layers. When thinking about the spatial distribution of D_b one of the primary drivers identified by the experts was land management practices (tillage etc.). This, however, was not present in the conceptual model as it the GIS data that would allow a spatial representation of land management practices was not available, which raises the issue of scale in relation to the application of expert knowledge. If the study was, for instance, conducted over a smaller area, it would be possible to identify and map more specific land management practices. This would allow the expert to provide more detailed insight into the variation within Land cover categories such as 'pasture' which was predominant within the study area. Alternatively, if the study was conducted at a national scale, climatic variables, such as temperature and rainfall, would become more influential predictors in the minds of the

experts. Differences in the aforementioned climatic variables were imperceptible to the experts at the scale of the study presented here. However, the group did raise the issue that there would be a notable difference in rainfall when comparing the west and east of Ireland, which could lead to notable differences in soil D_b.

4.4.5 Modelling Approach

It has been suggested that BNs may not be suited to modelling detailed spatial representations or making highly accurate predictions (Chen & Pollino, 2012). This chapter argues, however, that this disadvantage can be offset by the clarity of the modelling approach (the process by which predictions are made) and the ability to model future scenarios, in particular, under different Land cover or climate regimes. The ongoing integration of BNs and geographical information systems (GIS) offers to provide a platform for improving spatial predictions using BNs (Grêt-Regamey & Straub, 2006). While the predictive power of all the models examined here was relatively low, there are a number of benefits that can be drawn from the modelling process itself.

This is because expert knowledge elicitation within a BN modelling framework unveils the tacit choices and assumptions made by those involved in the production of (in this case) soil maps (Hodgkinson et al., 1999). This is of particular interest to the DSM community, where there is a drive to make use of the wealth of knowledge available within a rigorous, scientific framework (Scull et al., 2003). Defining the relationship between variables can be a very difficult task. Thus, even if the predictive power of a model is not particularly high, the development of a clear conceptual framework and a clarification of expert decision making can provide a stepping off point for future work. Moreover, the success of the model needs to be assessed in the light of its intended uses.

There should also be a distinction between the accuracy of the expert's knowledge and the accuracy of the elicitation process, as the two do not mean the same thing. An elicitation can be deemed a success if it accurately reflects the expert's beliefs about a situation, regardless of whether those beliefs are an accurate reflection of real life (Garthwaite et al., 2005).

The aim of this chapter was to determine the possibility of using expert knowledge in place of empirical data for the prediction of D_b at a landscape scale and, by proxy, to assess how well this approach could be applied to other soil properties. Regarding the application of BNs to other soil properties, one potential issue is that the knowledge required by the model is decided *a priori*, rather than assessing the available knowledge and assessing whether an expert-based modelling is a suitable approach. This is something of a cyclical argument as the limitations (or lack of them) of expert knowledge regarding a particular subject will not be discovered until the elicitation has been completed and the results of the model assessed. Low Choy et al. (2009) advocate a 'natural cycle of learning' where the findings of one model or study can become the starting point for the next. If expert elicitation using a particular technique exposes the limitations of either the technique or the available knowledge, it can direct future work towards a different approach, since there are often a number of ways that expert knowledge can be used in environmental modelling (O'Leary et al., 2008; Low Choy et al., 2009).

4.5 Populating a Soil Classification System with D_b Values

One of the key areas of investigation throughout this thesis concerns the choice between representing soil as a continuum on a gridded approximation of a continuum or as a set

of discrete polygons. As the primary use for D_b data is as an input parameter to other models, the focus has been on producing a gridded representation of the spatial variation. Despite this, soil classification and polygon-based maps are still a frequently adopted method of representing soil spatial variation, such as in the ISIS project. Gridded predictions of D_b have been produced using both data mining and expert knowledge so it is possible to assess whether these predictions can be utilised in a class-based soil mapping system. Of the three counties within the study area, County Waterford was the only one included in the original An Foras Taluntais soil survey (Gardiner & Radford, 1980). As such, a 1:100,000 scale soil series map was produced for County Waterford (Figure 4-8: Soil bulk density predictions by soil associations9) which included detailed profile descriptions for 32 of the series present, including reference D_b values (Diamond & Sills, 2011).

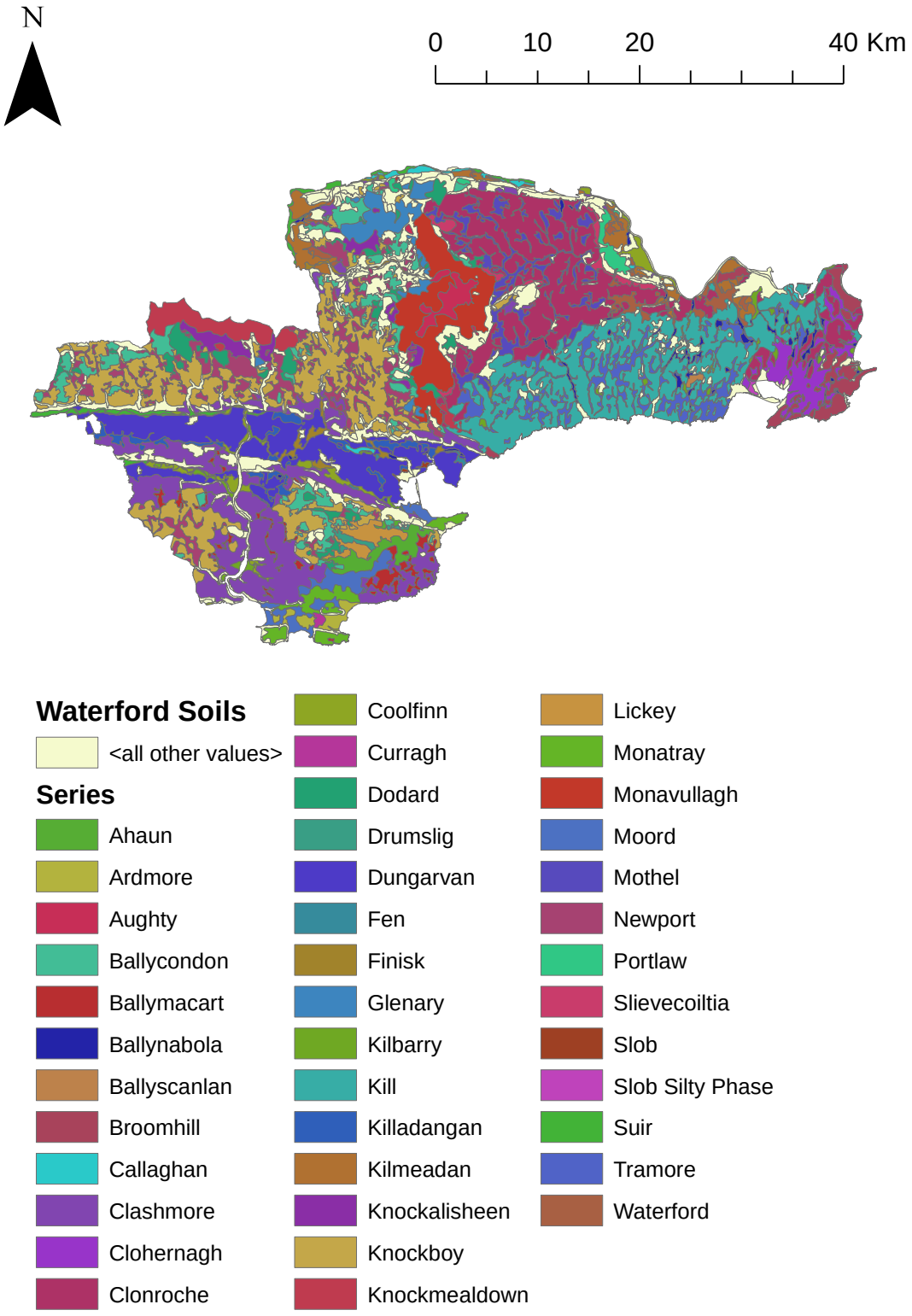


Figure 4-9: Soil series map of County Waterford

One of the criticisms of the polygon-based mapping approach is that it typically does not represent the within-class variation inherent for a given soil property (Heuvelink & Webster, 2001). By sampling the gridded predictions of D_b by series polygon using the 'Zonal Statistics' tool in ArcGIS (Beyer, 2004), it is possible to calculate a mean D_b value and 95% confidence interval limits for each of the RF, Naive BN and Hierarchical BN. These predictions can be compared to the reference D_b value, which is derived from measurement from a single soil pit to represent the typical D_b value of a given soil series (Diamond & Sills, 2011). For the majority of the soil series in Co. Waterford, the three predictive models broadly agree on D_b, as generally the 95% confidence intervals of each model overlap with the others (Figure 4-8). This is a significant finding, as the RF model was built solely using empirical data, while the naive and hierarchical BNs were both expert-derived. In particular, there is good agreement between the hierarchical BN and the RF model, which are better able to identify extreme D_b values, in particularly series with lower D_b. This suggests that it is possible to use expert knowledge as a proxy for empirical data.

The reference D_b measurements generally fall within the 95% confidence intervals of at least one the models. This trend would be less prevalent without the hierarchical model which has considerably larger 95% confidence intervals than the other two models. This can be attributed to its simplified structure and data requirements which lend a stepped appearance to the predictive map produced (Figure 4-7c). Although it still produces a continuous prediction, the D_b values appear to be divided into four distinct groups relating to the model structure. This has the benefit of improving accuracy when predicting low D_b values. However, if a series intersects two or more of these distinct groups, the 95% confidence interval surrounding the mean prediction becomes much

larger than that of the other models. This is not necessarily a negative, as can be seen from the relationship between the reference (measured) and predicted D_b values. As (Figure 4-8) shows, without the hierarchical model, many of the reference D_b values would fall outside the 95% confidence intervals of the models. This suggests that the large 95% confidence intervals associated with the Hierarchical model more accurately represent the within-series D_b variation.

There are five series whose indicative values fall outside the 95% confidence intervals of all predictive models; four of which are above the upper bounds (Ardmore, Clashmore, Curragh and Dungarvan) and one of which falls below the lower bounds (Clohernagh). As this is the exception, this prompts further investigation into possible explanations as to why none of the predictive models can capture these representative values. The information regarding the soil series is taken from the 'Soils of Co. Waterford' Soil Survey Bulletin (Diamond & Sills, 2011). For two of the series, Ardmore and Curragh, a possible explanation appears relatively straightforward. Both series represent only a very small land mass within Co. Waterford (0.38% and 0.07% respectively). This suggests that there may be insufficient data for the RF model to accurately predict accurately in these regions. With reference to the soil-landscape modelling approach, these soils occur so infrequently that it is difficult to establish a soil-landscape unit associated to these series'. Furthermore, the properties of the soils in these series' make them unusual and hence harder to predict. The Ardmore series is characterised by human modification over centuries, specifically the addition of sand and seaweed. This goes some way to explaining the high reference D_b value, as a high sand content is often associated with high D_b (Calhoun et al., 2001). The Curragh series, which is related to the Ardmore series, is also characterised by anthropogenic

disturbance. Furthermore, it is noted to have a typically low carbon content for Irish soil, which can explain a higher than expected D_b.

An explanation of the disparity between the reference and predicted D_b values for the Clashmore and Dungarvan series' (which cover 9.6% and 6.6% of the land surface respectively) is less obvious. The Clashmore series is classified as a 'Brown Earth (Typic Dystrudept)' meaning it describes a group of soils which are transitioning between brown earth and brown podzoilc groups. It is typified by below-average carbon concentrations, which suggest a higher than average D_b. However, surveyor's comments regarding the D_b in the series claim it will vary from moderate to high values (Diamond & Sills, 2011). On reflection, there are probably two main reasons that it falls outside the 95% confidence intervals. Firstly, none of the models appear to predict higher D_b values particularly well, as they are generally underestimated (Figure 4-5 and Figure 4-6). Secondly, the reference value for the Clashmore series is relatively high and it could well be the case that this is not representative of the series as a whole. Evidence to support this suggestion comes from the fact that there were two D_b samples taken from the Clashmore series other than the reference value, both of which fall within the confidence intervals of the RF and hierarchical BN models. Moreover, as a series which represents the transition from brown earth to brown podzolic, it is reasonable to expect areas of lover D_b which are typically associated with Podzolic soils. It is probable that the same two reasons (the failure of the models to predict high D_b soils and a particularly high reference value) can be applied to the Dungarvan series as well. However, there is no empirical evidence to support this assertion. The one soil series in which was the D_b was consistently overestimated was Clohernagh, which covers 1.89% of Co. Waterford. This can be explained by an unusually low reference value for a

mineral soil (0.79 g cm^{-3}). The series itself is described as poorly drained and as having high density, which is not reflected in the representative D_b value. The reason for this is that there are two A-horizons in this particular soil. The upper A horizon is only 10cm deep and has a low D_b value while the second A-horizon (A2) is much denser. In light of the soil description, it appears that using the A2 horizon D_b value would have been a more appropriate representation of this soil as a class.

One of the benefits of soil classification is the amount of information that the polygons represent. A corresponding criticism is that the information recorded in the polygon is tacit and hence not easily interpretable. A further criticism is that the within-class variation of soil properties is not represented. DSM approaches can offer a solution to some of these problems by combining the gridded and polygon approaches. In this way it is possible to populate a classification system that not only gives a reference D_b value (e.g. the model mean value) but also indicates the amount of in-class variability. As this is stored digitally, it makes using the information more straightforward. This method gives an approximation of the upper and lower bounds of the D_b values in a soil series without the need for a large amount of data and a stratified sampling scheme. In terms of using expert knowledge, this study suggests that it can be used as a proxy for measured data. Depending on the method used, the predictions can be very conservative, tending towards the mean, with corresponding narrow confidence intervals or more accurate predictions of extreme D_b values, with an associated increase in the size of the 95% confidence intervals. These large confidence intervals are not necessarily a negative, as they highlight the variability of some classes, which will tend to be underestimated by a RF model which has been trained on limited data. The investigation of the reference values which lay outside of the 95% confidence intervals

of all models highlighted the problem with using a single reference value for an entire series. When the reference value is at the extremes of the range (as was seen with the Clashmore series), it is unlikely that this will accurately reflect the average within-class D_b. This would be problematic when a spatial representation of D_b is required for stock assessment, for example, because it would introduce error without giving any indication of the magnitude of that error.

4.6 Conclusions

This study found that expert systems are able to identify the same broad spatial trends in D_b variation across the landscape as a RF data mining method. Using expert knowledge to populate the CPTs, as opposed to using data was, at best able to describe just over 40 percent of the variation in D_b. The Naive network had limited ability to predict D_b; it was able to explain around 25 percent of the variation but the vast majority of the predictions were around 1.1 g cm^{-3}. This is typically a 'medium' D_b value, meaning in essence, the naive network produced a very narrow range of medium value predictions. In particular, the naive network failed to identify soils with very low D_b values. This was principally ascribed to the method used to collate expert opinions; an opinion pool or group mean. Once aggregated, the conditional probabilities for each node were generally quite uncertain, meaning that often there was no clear relationship between variables. In the naive model, this was further complicated by the model structure itself. Many nodes, such as those regarding climatic variables, where the experts were highly uncertain had the same influence as those about which the experts were more certain of, such as Land cover and soil association. As a response to this, the hierarchical model; a much-simplified expert knowledge model, was used to identify large-scale spatial trends in D_b. The results of the independent validation and the similarity of predictions with

those generated by the RF model suggest that this was a successful, especially as the relationships between variables were expert-derived.

When populating the ISIS soil series classification, there was broad agreement between the RF and hierarchical BN models in terms of the trends in D_b across different series. Generally, the hierarchical BN was the model which had the largest confidence intervals. This meant that the pre-existing D_b reference values for the series typically fell within the confidence intervals of the hierarchical predictions, suggesting that this model was best suited to representing the scale of within-class variation of D_b Irish soils.

5 Integrated Discussion and Conclusions

This chapter summarises the project as a whole and discusses the results from the data-mining and expert knowledge derived models for both England and Ireland. By making reference to the literature, the aim of this chapter is to integrate the results into the wider discipline of digital soil mapping. Specifically, this chapter will discuss how the results address the aims of the study, how they contributed to knowledge and the wider implications of the findings.

5.1 Presenting the Problem

Increasing pressure on the Earth's resources has led to an increase in the demand for data about soil, which is necessary to "characterise the physicochemical, biological and hydrologic conditions of ecosystems across continents" (Grunwald et al., 2011). Many hydrological or climatic models (Harrison et al., 2008) require spatially-explicit, high-resolution digital data on specific soil attributes in a gridded format. This is at odds with the expert-derived, polygon-based, soil taxonomic class maps, which are the traditional method of representing the spatial variability of soil. The disparity between existing and required data is a complex issue for several reasons. While there is an increasing demand for data, there is contemporaneously a reduction in the amount of empirical data being collected, primarily due to financial constraints. Models which are typically used to create gridded predictions of soil properties (geostatistical models) require substantial amounts of sampled soil data compared to the amount of data typically used to produce a traditional soil classification map. This is because most existing soil classification maps have been produced during a soil survey, meaning empirical data was used in conjunction with expert knowledge, hence reducing the amount of sampling required. Unlike geostatistical models, however, expert-derived soil classification maps

rarely provide any measure uncertainty, nor represent the within-class variation of soil properties.

Digital soil mapping (DSM: e.g. McBratney et al., 2003) is a method which can make use of the strengths of both approaches to produce either gridded or polygon-based soil class maps. Unlike pure geostatistics, which treats variation in soil attributes as random, DSM is based on a "soil-landscape mapping" paradigm (Hudson, 1992) in which soil variation is determined by its relationship with a range of other environmental variables. By developing a statistical relationship between landscape and soil attributes at a limited number of sample points, it is possible to extrapolate across the landscape and to estimate the uncertainties associated with these predictions.

At the very largest scale, the demand for up-to-date, high resolution soils data is being driven by international policy makers, in reaction to global issues such as climate change, loss of biodiversity and the degradation of soil and water resources; which are all exacerbated by a rapidly increasing global population (National Research Council (US), 2001). Soils are also vital for agricultural food provision, for regulating global biogeochemical cycles and for supporting ecosystem services (Grunwald et al., 2011). Precision agriculture requires spatially explicit quantitative soils information for uses such as optimal application of fertilisers (Cassman, 1999). Data are also required to inform legislation used to protect soils, which have been deemed a non-renewable resource, vital for food production, nutrient cycling and water quality (Creamer et al., 2010). Empirically-produced DSM models have some potential for satisfying the demand for such data. However, there is also a need to better understand the spatial variation of soil properties, as part of a wider understanding of ecosystem structure and function (Grunwald, 2009). For this reason, models which can incorporate expert

knowledge relating the soil to the wider landscape are of interest, particularly as they can potentially reduce reliance on sampled data. More than two thirds of the Earth's soil remains unmapped (Nachtergaele & Van Ranst, 2003). The development of methods to better estimate soils data at a range of spatial scales and, particularly, in situations where measured data are scarce is, therefore, a key research topic.

The primary focus of this thesis is the spatial prediction of soil bulk density (D_b) which is defined as oven-dry mass per unit volume of a soil (IUSS 20 Working Group, 2006). It is of interest because it is typically measured at the point scale, but it is most frequently required at the landscape scale.

5.2 Literature Review Summary

The purpose of the literature review was to define the context of the work, examine the approaches taken by other comparable studies and to identify existing gaps in current knowledge. In summary, the literature review identified a range of different methods for representing the spatial variation of soil types and soil properties and highlighted the advantages and drawbacks of models used to produce spatial predictions about soils. The first stage of the literature review was to examine how the spatial soil variations were traditionally represented by soil surveys. Understanding the processes and limitations of the soil survey is important, as many of the maps produced using DSM techniques use 'legacy data' (in the form of soil classification maps and soil sample data used to validate those maps) as input data. The soil survey in the UK classified soils in the landscape based on homogenous soil properties (Avery, 1980). The boundaries between classes were delineated using a combination of empirical data, ancillary data (such as aerial photographs and geological maps) and expert knowledge.

The combination of technological advances (especially in computing), a desire for data detailing soil attributes rather than classes and issues regarding the tacit nature of the knowledge used in the traditional soil survey have prompted the development of digital soil mapping (DSM). Essentially, DSM is an attempt to maintain the soil-landscape paradigm used for soil classification, while overcoming the perceived limitations of traditional soil survey derived maps. DSM can be used to map either class or attribute using a set of reproducible statistical rules. In addition, it provides a measurement of the uncertainty associated with predictions. Two, non-parametric data-mining techniques were used in this thesis: artificial neural networks (ANN) and random forests (RF). These have both previously been used successfully to model a number of soil attributes (Behrens et al., 2005; Häring et al., 2012).

An issue concerning the use of data-mining methods for soil mapping and modelling relates to the interpretability of these models. Trends in DSM reflect a shift from models used purely for classification towards those which can provide or reflect better understanding of the spatial variability of soil (Grunwald, 2009). This introduces the topic of expert knowledge as a resource for DSM applications. Here, there is a clear link to the traditional method of soil survey, which is, in part, produced using a tacit expert system (the soil surveyor's knowledge and understanding). To be of use in a DSM context, expert knowledge must be applied within a quantitative framework. To this end, this study also considers the use of Bayesian Networks (BN). A BN allows knowledge to be structured in terms of probability, and has the benefit of being able to integrate expert knowledge and empirical data in an easy-to-interpret model.

5.2.1 Research Opportunities

The literature review identified several research opportunities which were used to develop the aims of the study. These opportunities are as follows:

- Quantification of the uncertainty in models of carbon stock assessment (and, by extension, other spatial models that require D_b as a parameter) due to the lack of a spatially explicit representation of D_b
- The use Bayesian Networks to model soil attributes and the introduction of expert knowledge as a potential method of improving the predictive accuracy of the models
- The use of expert knowledge as a stand-alone resource for digital soil mapping and the quantification of expert systems
- Investigating the difference in mapping soil attribute by taxonomic unit in comparison to using a continuous spatial prediction (for soil D_b)

5.2.2 Aims and Objectives

The aims of the thesis were as follows:

- To investigate the utility of soil-landscape models to produce spatially explicit maps of soil bulk density
- To demonstrate that expert knowledge can be used to improve soil-landscape models for the prediction of D_b

To achieve the aims, the following objectives were proposed:

- To develop a set of PTFs for soil D_b using a range of data-mining techniques developed from soil textural properties and organic carbon content and then to

attempt to improve predictive capabilities by including a range of soil-forming landscape attributes

- To test whether soil-forming landscape-scale variables alone can be used to predict D_b and, from this model, to produce a spatially explicit map of D_b
- To demonstrate the importance of a spatially explicit representation of D_b in relation to the development of soil carbon stock inventories.
- To test whether a Bayesian Network can be used as a suitable data mining tool to predict D_b
- To show that incorporating expert knowledge in the model structure of a Bayesian Network can improve the accuracies of prediction.
- To develop a naive Bayesian network to predict D_b using expert knowledge as a proxy for data
- To develop an expert structured Bayesian network to predict D_b using expert knowledge as a proxy for data
- To populate a soil taxonomic system with D_b values generated using data-mining and expert knowledge-based predictions and to compare the results to the reference values used for soil series

5.3 Discussion

To assess the outcomes of these objectives in relation to the thesis structure, the major findings, implications, limitations and contributions to knowledge will be discussed on a chapter-by-chapter basis.

5.3.1 Chapter 2: Modelling Soil Bulk Density at the Landscape Scale

5.3.1.1 Summary

The purpose of the work described in Chapter 2 was to develop a spatially explicit prediction of D_b. To accomplish this, the first step was to develop PTFs predicting soil bulk density using soil textural properties and OC content as predictors. The statistical methods used were multiple linear regression (MLR), which has often been used elsewhere (e.g. De Vos et al., 2005) and two non-parametric data-mining approaches: Artificial Neural Networks (ANN) and Random Forests (RF). The reason that these models were used was to facilitate the next stage in the modelling process: the explicit incorporation of a suite of ClORPT landscape attributes in models. Many previous studies predicting D_b improved their results by stratifying the data on the basis of Land cover or geology (e.g. Hallett et al., 1998; Calhoun et al., 2001; Steller et al., 2008). Here the purpose was to include landscape characteristics as predictors in an attempt to improve the accuracy of predictions. These data-mining approaches represent a shift from the semi-empirical soil survey methods of producing soil maps, towards a fully empirical method. One potential pitfall of this approach is that it is possible to "overfit" the data, meaning the model describes random error rather as well as the actual relationship between variables. To avoid this, the models produced in Chapter 2 were validated using independent data.

The first two steps were used to predict D_b at the point scale. In order to make predictions at the landscape scale, the predictor variables used need to be spatially explicit. Soil property variables such as soil texture properties or OC content cannot be used because they are not known beyond the point scale. The third set of soil-landscape models developed were used to create a spatially explicit map of D_b across the study

area. To illustrate why this was useful, the carbon stock based on the sampled OC content was derived for the study area using a mean and spatially-explicit estimations of D_b. The difference in both stock and associated error was used to show the value of this approach.

5.3.1.2 Key Findings

Chapter two produced a number of findings relevant to the creation of PTFs and soil-landscape models for the prediction of soil D_b. Of the models which used soil textural properties and OC content as predictors, which are analogous to the traditions PTF prediction of D_b (Rawls, 1983), the choice of statistical model can have a significant impact on the ability to predict D_b. This impact is horizon-dependent; in the A-horizon the non-parametric models (RF and ANN) significantly outperformed MLR, whereas, in the subsoil MLR and ANN were significantly more accurate predictors than the RF method. Overall, ANNs were the models which provided the highest predictive accuracy for both horizons.

The fact that the non-linear methods were best able to predict D_b in the A-horizon may be attributed to the fact that the A-horizon is the most the most susceptible to human influence. As this is the case, soils with similar textural properties and OC contents, may have quite different bulk densities due to the influence of, for example, Land cover and specific land management techniques (e.g. ploughing, rolling etc). Non-linear models are better equipped to represent this variation. In the subsoil, models were less able to describe the variation in D_b. Relative model performance also varied as MLR and ANN outperformed the RF model. Tree-based methods have had mixed results regarding the prediction of D_b and it has been suggested that regression tree models have limited predictive power when relating particle size distribution to D_b (e.g. Tranter

et al., 2007), whereas multiple additive regression trees, have been shown to produce significantly better results (albeit for a very different study area) (Martin et al., 2009).

Including the ClORPT landscape attributes in the models did not cause the uniform improvement in predictive accuracy which was expected. While it did improve predictions made by the RF and MLR models, reflecting the findings of other D_b studies (Hallett et al., 1998; Martin et al., 2009), the predictive powers of the ANN model decreased. A decrease in the predictive power of ANN models following the addition of predictor variables has been attributed to the inclusion of predictors which have a weak correlation with the variable being predicted (Amini et al., 2005). This can lead to 'overfitting' the model (Tranter et al., 2007). Indeed, including even a single variable which has a low correlation with D_b will decrease the predictive accuracy of the output (Keshavarzi et al., 2010).

The most crucial finding in Chapter 2 was the ability of the ANN and RF models to predict over 50 percent of the variation in D_b using landscape variables alone. This facilitated the production of a spatially explicit gridded prediction of D_b. This is particularly important as the lack of a spatially explicit representation of D_b is one of the primary sources of error in models predicting soil carbon stocks (Grimm et al., 2008). To illustrate the impact of this model, the gridded D_b estimation was used to estimate a hypothetical carbon stock, which was compared to the stock calculated by the weighted average of the samples in the region. The results showed that the two methods produce notably different accuracies in stock estimations, despite soil carbon concentrations being the same for both models. There were also significant regional differences in stock estimation themselves, in some regions the difference in carbon stock was as much as 15 tonnes per hectare. One point to address is that irrespective of the model

used, nearly half the variation in D_b remained unexplained. This can be attributed to the input data and the various model structures. D_b was sampled at a point scale, however, many of the processes which can influence D_b at the point scale were not represented by the input data. This is due to both the coarse resolution if the raster datasets used and a lack of data representing, for example land management. Moreover, as the relationship between the landscape and D_b is not well defined, the predictor variables used might not have been those beast suited to explaining landscape scale variation in D_b. The statistical models themselves are also a source of uncertainty, as the models make unknown simplifications and assumptions to produce predictions. One of the problems with using black box modelling techniques is that it is difficult to ascertain whether the models are making predictions which make pedogenic sense.

5.3.1.3 Contribution to Knowledge

The addition of ClORPT landscape variables (to soil properties) as predictors of D_b in PTFs resulted in improved accuracy. However, this was both model- and horizon-dependent. Specifically, prediction of D_b in the A-horizon and in the subsoil will be improved by including landscape variables when using a RF or MLR model. The addition of these variables in an ANN model can be detrimental to model performance.

Creating a spatially explicit gridded prediction of soil D_b using non-linear, non-parametric models is a novel approach to mapping the spatial variation in soil D_b and directly addresses a gap in existing knowledge (Grimm et al., 2008). The value of this technique was demonstrated by estimating soil carbon stocks for the study area, including associated error.

5.3.1.4 Implications

The implications for the creation of new PTFs for D_b are that landscape variables have the potential to improve predictions but require carful testing and a correct modelling framework. If landscape variables are not available, using a non-linear, non-parametric model such as RF or ANN will give more accurate predictions in comparison to MLR models. While this and other studies suggest that adding landscape variables should improve tree-based models used to predict D_b (Tranter et al., 2007; Martin et al., 2008), it is necessary to evaluate carefully which variables are included in ANN models, as the addition of extraneous variables can result in diminished predictive capabilities.

The ability to predict D_b as a gridded surface has wider implications, which are relevant to a range of stakeholders. As a response to the growing demand for high resolution soil data required to address a number of environmental issues; the GlobalSoilMap.net project (Sanchez et al., 2009) aims to produce a global map of soil properties, including D_b. The gridded model of D_b can potentially become the basis for mapping this property on a global scale. There are also specific implications for modelling D_b using this method. Quantifying uncertainty in the carbon cycle is important for modelling climate change on a global scale. A large but particularly uncertain fraction of the terrestrial carbon stock is the carbon stored in soil (Matthews et al., 2000; Throop et al., 2011). Since the gridded method of D_b prediction can improve carbon stock estimation and error propagation, it could be used to improve models linking terrestrial and atmospheric carbon stocks (e.g. King et al., 2007). This approach to representing the spatial variation across a landscape can also be applied to precision agriculture (Cassman, 1999) and a variety of hydrological models (Wösten et al., 2001)

5.3.1.5 Limitations and Potential Improvements

One of the limitations of the study presented in Chapter 2 is that the models developed are 'black-box' and hence cannot easily be interpreted in physical terms. Independent validation of the models limited the potential problems of overfitting. However, there are other issues concerning these data-mining approaches. As it is not clear which processes the models are representing, it is not advisable to apply these models to other regions. Moreover, it is difficult to learn from the modelling process. Other than ranking which variables are most important for the models' predictive capabilities, the ANN and RF approaches have provided limited insight into the relationship between soil D_b and the wider landscape. Chapters 3 and 4 attempted to address these issues.

A more specific limitation of the approach used in Chapter 2 concerns the choice of input datasets. Due to differences in both pedogenesis and location-dependent data availability, there is no definitive set of environmental variables used for DSM, let alone the spatial prediction of D_b. For instance, in terms of topography, while elevation, aspect and slope are commonly used, wetness index and distance to stream have also been employed. The study could have been improved by a more rigorous method of testing and selection of input datasets and, possibly, by employing spectral reflectance, or other remotely sensed data, such as those gathered by radiometric techniques, as predictors (Moreira et al., 2009).

A further limitation relates to how the topographic input datasets were derived. The basis for the predictors such as slope, elevation and soil wetness index were derived from an original 10 m resolution digital elevation model (DEM). In DSM applications, using the most fine scale spatial data available is common practice, but this has does not always lead to the most accurate prediction of soil class. This is particularly true of

relatively flat, homogenous landscapes (which represent the majority of the Midlands study area), in which predictive accuracy is improved by using a DEM at a coarser resolution (Cavazzi et al., 2013). The study could have benefitted from empirical testing of the topographic data layers used as predictors.

One of the problems with using legacy data relates to the discrepancies between the dates on which each set of data was collected. The soil survey data were collected between 1970-1987, the climatic data were averaged from records spanning 1970-2000 and the other datasets were compiled in the intervening years. Combining data from a number of sources, collected on a number of different dates will inevitably introduce some error to predictions.

5.3.1.6 Future Work

The prediction of D_b using a number of different linear and non-linear modelling approaches has shown that model performance will vary depending on the soil horizon and predictor variables used. Performance will also vary spatially with each model predicting more or less accurately at different locations. As this is the case, one of the questions raised is whether an ensemble approach to modelling could improve predictions. The concept of this method is to improve predictive accuracy and reduce uncertainty by integrating model predictions, with weighting towards those models which most accurately predict, given a specific set of circumstances (Lee et al., 2012). To expand on this point, a proposed reason for the spatial variation in model performance is that the different models represent processes acting at different scales. In some areas, large scale variation in climate may be the primary driver of variation in $D_{b,}$ whereas in others, local variation in topography might be more important. Clearly, it is difficult for a single model to represent all the scale dependent processes interacting

at a given location. For this reason, one alternative modelling approach would be to use a series of nested models within a Bayesian decision making framework, where a different data-mining model is deployed to make predictions based on what is deemed to be the primary driver of D_b variation at a given location. Regarding the creation of a digital soil map for the world (Sanchez et al., 2009), the project aims to have predictions of soil properties on a 3D grid. This would require the incorporation of a continuous change of properties with depth, as opposed to modelling a separate gridded surface for each horizon. One possible avenue for investigation would be to use the gridded model for topsoil in combination with a depth function (Veronesi et al., 2012), in order to model D_b in three dimensions.

Chapter 2 demonstrated the potential improvement in carbon stock assessment made possible through the use of a gridded estimate of D_b. This was, however, only a hypothetical stock assessment as the carbon concentration was kept constant. The next stage of this work would be to combine the gridded prediction of D_b with a gridded prediction of OC concentration to produce a more realistic estimate of OC stock at a national scale. The findings of such an exercise could be compared to other national scale estimates (Smith et al., 2006) and subsequently be applied to large scale studies regarding the size of terrestrial carbon stores (e.g. Bellamy et al., 2005). Beyond carbon stock calculation, the gridded model of D_b could also be applied to a range of hydrological models (e.g. Miller & White, 1998).

One continuing gap in knowledge relates to the contradictory assertion that the accuracy of ANNs are unaffected by extraneous predictors (Behrens et al., 2005) and the findings of this study and others (e.g. Amini et al., 2005) that ANNs are prone to overfitting and hence are negatively impacted by the addition of variables which are not strongly

correlated to that which is being predicted. This may well relate to how the parameters of the network are set during training. However, this is an area which would benefit from the development of clearer guidelines.

5.3.2 Chapter 3: Using Bayesian Networks for Digital Soil Mapping

5.3.2.1 Summary

One of the limitations of the findings of Chapter 2 was that the models used to create the spatially explicit predictions of D_b were black box. Consequently, it was difficult to interpret the physical processes and relationships underpinning the predictions made. To resolve this, the potential of using a Bayesian Network to produce a spatially explicit map of D_b was explored in Chapter 3. This chapter consists of three stages. Firstly, a naive Bayesian network was used to make predictions of D_b using the landscape predictors identified in the Chapter 2. However, this model does not accurately represent the interactions between variables. Hence a second BN model was created using an expert-defined structure. This model was an initial attempt to include expert knowledge in the prediction of soil bulk density. In this case, expert knowledge was used to produce a model that more accurately reflects the landscape process affecting D_b. The final section of Chapter 3 compares the modelling results and mapped output with those of the data-mining approaches in order to ascertain whether BNs can feasibly be used to predict a continuous soil attribute. Assessing the predictive accuracy of a BN is itself is something of a novelty, as it generally happens so infrequently (Aguilera et al., 2011).

5.3.2.2 Key Findings

The key findings of Chapter 3 are that BNs can be used to create PTFs for D_b and can produce predictions with comparable levels of accuracy to those of MLR, RF and

ANNs. That said, it should also be noted that the BN-derived predictions are generally slightly poorer than the best performing data-mining techniques in terms of predictive accuracy, especially for the subsoil horizon.

In terms of creating the gridded prediction of D_b, the optimised BN have very similar predictive capabilities to the RF and ANN model with the benefit of being a more easily interpretable model.

Contrary to expectations, the addition of expert knowledge, in the form of an expert-derived model structure did not improve the predictive power of the model. This can primarily be attributed to the experts disregarding the direct affect many environmental covariates have on D_b. In the expert-derived model, only land use and soil association were assumed to have a direct affect on D_b, whereas the naive models, represents the influence of many more environmental variables. This suggest the expert models require further iterations to accurately capture the processes contrioling the variation of D_b in the landscape. The expert-derived model had a similar predictive performance to the naive network and substantially less than the optimised naive network.

5.3.2.3 Contribution to Knowledge

Bayesian networks have been applied successfully to the prediction of soil classes (Mayr et al, 2008; Mayr & Palmer, 2006) but have seldom been applied to predict continuous properties. They have not previously been applied to the landscape scale prediction of D_b. The work presented in Chapter 3 shows that BNs can and should be considered for this application, along with the data mining techniques used in the previous chapter. In comparison to other models which used a combination of Bayesian statistics and expert knowledge to predict soil attributes (of which there are very few)

(e.g. Corner et al., 2002) the BN method presented has the important advantage of not assuming that the predictor variables are independent. The expert structure-derived model explicitly represents the interactions between variables rather than relying on an unrealistic assumption of independence.

5.3.2.4 Implications

The major advantage of modelling the relationship between soil and landscape variables using a BN is the manner in which the plausibility of the model can be assessed (Finke et al., 2012). In black-box models, this is limited to examining the model inputs and outputs without examining the process by which predictions are made. The clarity of the BN modelling approach is superior to data-mining approaches in that it can be implemented from two unique standpoints:

(1) From a purely data-mining approach, if little is known about the system under investigation, the relationships between variables generated by the BN can be used to identify trends between variables. The major difference between this and, for instance a CART approach, which can be used generate a series of rules to divide up the landscape, is how the BN deals with uncertainty. The BN model explicitly quantifies the uncertainty inherent in the data, whereas the CART model simply produces a set of definitive rules. While getting clearly defined rules may intuitively seem more desirable, it can be problematic since CART models are very sensitive to changes in the data and notably different rules can be derived depending on which variables are included within the model (Scull et al., 2003).

(2) If there is a greater understanding of how the variables in the system interact then it is possible to structure the model *a priori* in such a way as to reflect interactions

between variables. As BNs are graphical models and hence easy to interpret, they are open to scrutiny and review from users or other experts and can be amended to reflect changes in process understanding. Moreover, both the naive and expert models can produce associated maps of uncertainty, meaning that map users can see how the model performs spatially. This is useful, as even highly uncertain maps can be helpful to decision makers (Carré et al., 2007).

5.3.2.5 Limitations and Potential Improvements

Chapter 3 uses BNs in an attempt to get more insight into the relationship between soil and the landscape. The work presented suggests that BNs can be used to predict continuous soil properties and that interpreting how predictions are made can be more straightforward. Using this approach, however, did not reduce the data-requirements necessary for the modelling process. The BN still relied on a significant amount of data for training and, subsequently, for validation as well as placing demands on experts' time to produce a model structure. That said, the full potential of expert knowledge, which can be used to determine the effect variables have on one another as well as determining model structure, was not utilised (Corner et al., 2002). If expert knowledge was used to determine conditional probabilities as well as the model structure, the empirical-data requirements during the modelling process would be reduced. This issue is addressed in Chapter 4.

A further issue with the study is that, contrary to expectations, the implementation of the expert structure within the model did not improve predictions. This expectation stemmed from the fact that a naive network (where all nodes influenced D_b equally) was clearly an unrealistic representation of the physical relationships which exist in reality. The purpose of using an expert-structured model was to more accurately represent the

interactions between variables. The failure of this approach to yield more accurate results was attributed to an underestimation of the number of environmental variables which had a direct impact on D_b (the nodes are directly linked to D_b). In the expert model, only Land cover and soil association are linked to D_b. However, sensitivity analyses of both the naive network and the data-mining approaches suggest that parent material and climatic factors can both have a significant influence. Another limitation is manifest in the input datasets available for the study. There are no GIS data layers that represent land management practices spatially, meaning they cannot easily be mapped. While the likelihood is that this would also improve the predictive powers of the data-mining models, it is of particular interest in the expert-derived models, as experts tend to think in terms of land management. One possible approach to mitigate this is to create intermediate nodes which amalgamate existing data to represent land management practices, if such data can be obtained (Aalders et al., 2011), although this will increase the uncertainty in the final mapped output.

5.3.2.6 Future Work

Since the experts appeared to underestimate the direct influence of climate (and the many derivatives of climatic indices) there is scope for a large-scale data-mining study into the effects of climate on soil class and property. A similar study might test the influence of topographic derivatives using ANNs (Behrens et al., 2005). Studies of this nature would need to be conducted on a large scale; ideally, national or continental, meaning the data requirements would be considerable. Using a large dataset would also enable the structure of the BN models to be determined using a data-mining algorithm such as the TAN learning structure, which is comparatively data intensive (Friedman et al., 1997).

Another area for investigation regards the methods by which models are evaluated. As shown in both Chapters 2 and 3, there are usually a number of methodologically suitable models which can be used to perform the same task. The question is how to choose between them or to decide which one is best? Generally, residual based tests (measuring the difference between observed and predicted values) are used. However, the residual method of model comparison is not definitive, as the models will, inevitability, be calibrated or validated on a limited dataset. Furthermore, different models will often provide similar measures of accuracy (as is the case with D_b). For these reasons, model performance using information theory, which assesses the information content of the models (Akaike; 1973; Pachepsky et al., 2006) might hold promise.

5.3.3 Chapter 4: The Application of Expert Knowledge to Bayesian Networks

5.3.3.1 Summary

Chapters 2 and 3 investigated soil D_b in a relatively data-rich setting in the UK. Chapter 4 builds upon the findings of the previous two in a comparatively data-poor study area (The Republic of Ireland). The purpose of this chapter was to investigate expert knowledge as a resource. To do this, a further three models were produced. An RF model was built and cross-validated using the limited data available, and from this model, a map of the spatial distribution of D_b was produced. As this method has been shown to work well (Chapter 2), it was taken as the benchmark for comparing the maps produced by expert knowledge. The second model built was a naive BN, where the conditional probabilities defining the relationship between D_b and landscape variables (the CPTs), were derived from expert knowledge rather than from data. The third model

was an expert-derived BN which used a hierarchical structure. The two 'expert models' were evaluated using the available data. The RF model was cross-validated using the same data used to train the model. As such, the modelling results between expert and RF models should not be compared directly, although the cross-validation is gives a strong indication of how the RF model would perform on an independent dataset. This method did, however, allow for a visual comparison of the expert and data derived maps.

The impetus for this study came from the Irish Soil Information System (ISIS) project which aims to complete a 1:250,000 scale national soils map for Ireland in order to conform to E.U. legislation and support ongoing soils research (Daly & Fealy, 2007). This project uses a range of data-mining approaches to produce a polygon-based soil classification map. This thesis makes an explicit contribution to ISIS by investigating the possibility of populating this classification map with D_b data. Using both expert knowledge and data-mining approaches, it was possible associate D_b values with soil series providing a mean value and a 95% confidence interval. This could, then, be compared with pre-existing reference values for each series.

5.3.3.2 Key Findings

The key findings of Chapter 4 were that expert systems were able to identify the same broad spatial trends in D_b variation across the landscape as an RF data mining method. Using expert knowledge to populate the CPTs, as opposed to using data, was, at best, able to describe just over 40 percent of the variation in D_b. The naive network had limited ability to predict D_b; it was able to explain around 25 percent of the variation but the vast majority of the predictions were around 1.1 g cm^{-3}. This is typically a 'medium' D_b value, meaning in essence, the naive network produced a very narrow range of

medium value predictions. In particular, the naive network failed to identify soils with very low D_b values. This was ascribed to the method used to collate the experts' opinions; an opinion pool or group mean. Once aggregated, the conditional probabilities for each node were generally quite uncertain, meaning that often there was no clear relationship between variables. In the naive model, this was further complicated by the model structure itself. Many nodes, such as those regarding climatic variables, where the experts were highly uncertain, had the same influence as those for which the experts were more certain, such as Land cover and soil association. As a response to this, the hierarchical model; a much-simplified expert knowledge model, was used to identify large-scale spatial trends in D_b. An independent validation, together with similarity with the quality of the RF prediction suggest that this was successful, especially as the relationships between variables were expert derived.

When populating the ISIS soil series classification, there was broad agreement between the RF and the hierarchical BN models in terms of the trends in D_b across soil series. Generally, the hierarchical BN was the model which had the largest confidence intervals. This meant that, generally, the pre-existing D_b reference values for the series typically fell within the confidence intervals of the hierarchical predictions. This indicates that this was the model best-suited to representing the scale of within-class variation of D_b for ISIS soil series.

5.3.3.3 Contribution to Knowledge

There has been a longstanding drive to formalise the inclusion of expert knowledge in the soil modelling process (Dale et al., 1989; Shi et al., 2009). Of the available 'expert-systems' approaches to soil modelling, Bayesian methods have been identified as one of the strongest candidate techniques for structuring expert knowledge (Skidmore et al.,

1996; Bui et al., 1999; Corner et al., 2002; Farewell, 2010). The expert system proposed in Chapter 4 differs from those in previous studies and is novel in a number of ways. Firstly, it uses a Bayesian network to predict a continuous soil property. Secondly, the relationships between variables are completely defined by experts and, hence, can be independently validated, giving a clear indication of the accuracy of expert knowledge for predicting the spatial variation of D_b. Thirdly, the hierarchical model is a unique an attempt to combine an expert-derived structure (representing interactions between variables) with expert derived conditional probabilities to predict a continuous soil property, in a single model.

5.3.3.4 Implications

Digital soil mapping is particularly challenging in countries with limited quantitative data. Many tropical countries, for example, lack empirical data but have an abundance of qualitative information in the form of soil surveys and classification studies. This is essentially a repository of expert knowledge, and information recorded in these surveys has been used, in combination with limited quantitative data, to form rules for predictive soil mapping, with limited success (e.g. Stoorvogel et al., 2009). Furthermore, the extent to which DSM approaches can be used in areas with very low soil sample densities as the basis of reconnaissance soil surveys has also been debated (Mora-Vallejo et al., 2008). The BN models developed in Chapter 4 and the hierarchical model, in particular, offer an alternative to mapping using knowledge-based rules derived from existing literature or geostatistical methods developed from limited data. In Chapter 4 knowledge-based models based on a group of experts rather than recorded data was shown to identify spatial trends in D_b variation reasonably well (when compared to a data-mining approach). The caveat must, of course, be made that expert knowledge is a

resource just like empirical data. If it is unavailable, then the model may not be applicable to regions for which there is a lack of expertise. This, however, is an area that requires further study regarding how adaptable purely expert-knowledge-based models can be. It may be the case that with limited 'calibration' to local soil conditions, experts may be able to adapt their knowledge and process understanding accordingly.

5.3.3.5 Limitations and Potential Improvements

The elicitation process and the results of Chapter 4 have identified some areas which could be improved. The failure of the naive network model to identify areas of low D_b and subsequent improvement in performance via the hierarchical model suggests that it may have been useful to develop separate models for organic and non-organic soils. The landscape variables within the hierarchical model were classified to reflect the primary relationship between D_b and the state of the variable in question. As such, the hierarchical model was a significant simplification of the relationships generated in the naive network. The primary reason for the improvement in the predictive power of such a simplified model, can be attributed to the fact that it was better able to identify the very low D_b values associated with peat and other organic soils. Clearly, it is difficult to model the complex number of interactions affecting mineral soils. The same models will probably not be applicable to organic soils, which (irrespective of other landscape conditions) will have very low D_b values. This issue was not addressed in the models developed on UK soils as there were few organic soils in the study area examined. By excluding the organic soils from the model, the overall predictive accuracy of the naive model may have been improved. Furthermore, the comparatively good result of the hierarchical method was achieved after including feedback from the original model. In this way, the weaknesses of the naive structure were deliberately amended. Feedback is

a crucial part of the elicitation process and the improved performance of the hierarchical model reflects this.

The resolution of the input data, in particularly the generalised soil map, was as was problematic during the elicitation process. The soils data came from a 1:575,000 scale soil map showing soil variation at the Great Group level. Within a Great Group, there is a lot of variation between soils, hence the experts found it hard to reflect the uncertainty inherent while still identifying between-class trends in D_b. It is possible that the high degree of uncertainty associated with the soil class D_b contributed to the poor performance of the naive BN. This shows clear parallels between the expert and data-mining modelling approaches. It is possible that both the expert and the Random Forest models would have performed better given a more detailed soil map. The variation present in low-resolution data may lead to greater model uncertainty for both expert and data-mining approaches.

A further limitation of the study relates to the use of the hierarchical model as an expert system. One of the primary drivers of innovative modelling techniques is the need for better understanding of the processes driving the spatial variation in soil properties (Grunwald, 2009). It is questionable whether the hierarchical model does this; it provides a framework for the experts to identify spatial patterns of D_b variation, but it has not necessarily helped the development of a greater understanding of the processes at work. The model is functional but possible has too many simplifications to improve process understanding and, in this respect, the naive network is probably a better model overall.

5.3.3.6 Future Work

Creating an expert structured model where the conditional probabilities were also expert-derived was considered and rejected due to the complexity of the conceptual model which would be needed. The number of joint probabilities required was deemed to be prohibitively large. However, the results of the hierarchal model suggest that this might actually be an approach worth pursuing. Many of the joint probabilities in the hierarchal model were deemed to be similar for different combinations of variables, reflecting the influence of a dominant variable. While this may weight probabilities in favour of soil class (depending on the experts used) it would allow the experts to highlight the relationships between specific classes more effectively. The use of expert knowledge within a Bayesian modelling framework is yet to be applied to complex soil models (Finke, 2012).

As mentioned above, there have been previous attempts to incorporate the uncertainty inherent in different GIS data layers within a Bayesian framework (e.g. Corner et al., 2002). Expert assessment of map purity is a potential subject for research since the feasibility of allowing experts to make these judgements, which experts to ask, quantifying their accuracy and how best to incorporate the spatial changes in uncertainty associated with proximity to class boundaries are all questions which have yet to be adequately addressed.

5.4 Reflections

In DSM, the process between identifying a problem to be solved and the creation of a final map (in this case of D_b) consists of a series of choices which will affect the outcome of the model and introduce knowledge to the modelling process. The formation of soil and the spatial variation present in its properties must obey certain fundamental

physical laws. Hence, if enough data could be recorded and the processes controlling variation were understood perfectly then the variation would certainly be deterministic. As this is not the case, there is a choice between treating the variation as stochastic or deterministic. In making this choice, expert knowledge has already entered the modelling process since there will normally be numerous potential approaches to solving the same problem. This study has adopted a deterministic approach to soil property modelling in the form of soil-landscape models and expert-knowledge-derived Bayesian Networks. Within a deterministic modelling framework, models can either be knowledge- or data- driven. Generally, DSM uses data driven approaches to produce predictive maps. However, in reality, expert knowledge is seldom completely excluded from the modelling process. By using a soil-landscape model, expert knowledge enters the model implicitly in the choice of datasets representing the environmental covariates; in particular, pre-existing soil maps.

This can even be the case for geostatistical models, which are frequently stratified on the basis of a number of landscape variables or even soil classes. This has been shown for D_b (which is rarely predicted using geostatistics due to the lack of available data) where the accuracy of kriging results are improved by including stratification on the basis of soil maps (Utset et al., 2000). As the inclusion of these covariates has the potential to improve the predictive power of data-driven models, it is clear that expert knowledge is an avenue for investigation.

The failure of the expert structured model to improve predictions in Chapter 3 and the limited predictive power of the expert models in Chapter 4 raises questions over the usefulness of the method. On reflection, D_b was a difficult soil property on which to base an expert system as it tends not to be considered at large spatial scales in the way

that soil class would be. Indeed an expert system would probably not be considered for DSM if a project was not subject to financial and time constraints, which can affect the ability to generate empirical data. However, this situation is rarely if ever the case and so, in situations where there is a shortage of data, expert systems have proved to be useful. This was demonstrated in Chapter 4, where expert systems were shown to be able to represent the pattern of D_b variation created by a data-mining tool with similar levels of accuracy. Moreover, as an alternative to a taxonomic representation of soil attributes (a single D_b value attributed to each series) they have the advantage of representing the variability of the property in question without the need for more sampling.

5.4.1 Viability of Techniques

This study has developed and tested a number of models, ranging from purely empirical (PTFs and data-mining), through semi-empirical (expert structured BNs) to expert systems. All of these models (with the exception of PTFs) were used to create a gridded prediction of D_b to satisfy the demand for a spatially explicit representation of D_b (Finke, 2012). The choice of which model to use depends on where this demand comes from and what data are available. If there are sufficient data and the spatial estimate of D_b is required as an input to another model, for example to calculate soil carbon stock (Grimm et al., 2008), then a data-mining approach is the preferred option. This is because this approach gives a clear measure of uncertainty associated with the prediction, which can be propagated through the model. Consequently this approach is recommended for the calculation of soil stock estimates with the caveats that models should be validated (independently validated where possible) in order to prevent overfitting and subsequent underestimation of uncertainty. Furthermore, careful

consideration of which environmental variables are included as predictors is required to produce models with the greatest predictive accuracy (it should be noted that this is true for all modelling approaches). The fact that there is broad agreement between modelled predictions of the spatial variation in D_b adds weight to the validity of the approaches. For the UK case studies in Chapter 2 and Chapter 3, both the data-mining and expert based models identified the same regions of high and low D_b. This suggests that non-parametric data-mining techniques can be used to identify trends in D_b, as can expert knowledge derived models. For most applications, knowing broad spatial trends should be sufficient. As data-mining techniques develop and the processes which drive the variation in D_b become better understood, the models will improve. Any discrepancy between predictions from each model can be attributed to differences in model structure. Inevitably, each model will capture slightly different processes or weight the influence of a given predictor variable differently. Explaining how each model is making predictions is not straightforward as the process is not easy to decipher in a black box model. One method would be to conduct some analysis of the residuals (of predictions) to see whether there is any pattern to where the models perform well or poorly. Understanding how predictions are made by a BN model should be more straightforward, as the relationship between variables is explicit. In this instance, further investigation by the expert is required to analyse what could be causing the model to perform poorly in places. It is probable that the resolution of the input data is insufficient to capture much more than half the variation in D_b, although this suggestion requires further study.

If there are sufficient data and the models are to be used as a form of knowledge discovery (Bui et al., 2006), then a semi-empirical, expert-structure BN might be

considered. The data-mining approaches provide limited insight into the interactions between variables, whereas by using the BN approach, the relationships between the soil and landscape are more readily interpreted. Unlike classification or regression trees (a data-mining technique often used for knowledge discovery), which provide a clear set of rules linking soil and landscape, the BNs will only identify general trends. While this may intuitively seem less desirable, the BNs are less susceptible to variations depending on what data is included in the model, and hence the trends indentified are more applicable to the wider landscape. The fact that an expert-structured model was not able to improve predictions suggests that, for D_b, the direct drivers of variation are not well understood. For the purposes of knowledge discovery, it is recommended that the modelling process be iterative, giving experts an opportunity to revise their opinion of model structure in response to previous and emerging results. An expert approach to modelling requires careful consideration of what expertise is available, for the prediction of landscape-scale D_b. In general, there appears, thus far, to be a lack of knowledge regarding the direct drivers of variation.

The thesis explored the possibility of using expert knowledge in place of empirical data in areas that are not data-rich. The results showed that this approach provided a viable alternative, implying that expert knowledge can be considered a resource for the creation of a gridded, as well as polygon-based, prediction of soil properties. While this finding has significant potential for mapping soils in areas where there are little empirical data (Hansen et al., 2009), it is important to consider how the reliability of these models is judged. Confidence in the results comes from comparisons with measured data in both the reference values for D_b for each series and the comparison between the predicted results generated using the RF model. Moreover, the results of

the expert knowledge models are evaluated by the model's ability to predict measured data. It is questionable whether predictions from a purely expert model could be used for the purposed of stock assessment, for example, without some form of quantitative assessment to evaluate the uncertainty associated with predictions. A more likely use for this technique would be to inform sampling schemes for future surveys (McKenzie & Ryan, 1999), or land management applications, such as identifying areas likely to be at risk of soil erosion (Aalders et al., 2011). A key consideration for the expert system is its interpretability. This is especially critical if empirical data are not available. An expert model needs to be open to review and amendment from other experts. This is a critical difference between the expert-based and data-mining models. The latter do not invite expert assessment regarding the processes that lead to predictions. The only way for an expert to evaluate a data-mining model, is to examine the output. For this reason, although the primary method of soil mapping is shifting from expert-based polygon delineation to gridded predictions of soil variation, soil polygon maps provide a rich source of data which can be used to develop and test new mapping techniques (Hartemink et al., 2012).

5.4.2 Landscape Scale Prediction

Landscape scale prediction of soil properties on a grid is typically produced using data-driven empirical models such as those generated from geostatistics or data-mining methods. The data-mining methods used in this study cannot be described as process-based models, as it is not clear which processes the models represent. The BN models can be described as semi-process-based as it is possible to infer the processes being represented from the relationships between variables. Landscape-scale process-based models are very complex, as the scale of the model inputs must correspond to the scale

of the processes being considered (Pennock & Veldkamp, 2006). The issues regarding scale add a layer of complexity to process-based modelling. Specifically, methods used to address the interactions between processes at different scales are an area of ongoing research for DSM applications (Cavazzi et al., 2013). The reason that process-based models are of interest is that they are based on an understanding of the environment. This means that, if there are areas in which process-models do not make accurate predictions, such limitations can be investigated in terms of landscape processes. If it transpires that the processes assumed to hold true across a study area actually vary depending on landscape conditions, then these models can be used for the purpose of knowledge discovery (Bui et al., 2006).

At present, data-mining methods are those best suited to describe the landscape scale variation in soil properties. However, the results from the BN models suggest there is scope to create more interpretable models based on current understanding of the processes occurring in the landscape. Given the limited number of model inputs and their relatively simple model structure, the BNs were able to identify trends in the variation of D_b, which were adequate for the data requirements of some land management practices and to inform sampling schemes for future soil survey data collection. Given that the processes which govern soil formation, and hence variation in properties, must obey certain physical laws (Webster, 2000) it is likely that technological advances in computer modelling techniques and data collection (remote sensing) will enable more accurate representations of the processes governing soil variation in future, and hence will better-enable landscape-scale process modelling of soil attributes.

In comparison to representing soils as a set of polygons, with uniform attributes derived from representative profiles, a gridded representation has been shown to be a generally preferable method. Even if the models used make highly uncertain predictions, which is the case for both the data-mining and BN models, an important improvement is that this uncertainty is explicitly acknowledged and represented. If soils data remain polygon-based, the lack of variability within a class will always be a limiting factor in physicochemical, biological and hydrological models for which the data are required. By incorporating a measure of uncertainty, this DSM approach to soil mapping can conform with the soil-landscape paradigm adopted by traditional soil mapping approaches (Hudson, 1992), while producing data to meet the requirements of modern users. If a soil classification map is required, producing a gridded model of soil attributes can provide a representation of the within-class variation as opposed to the representative profile-derived single measurement. The potential of this method is demonstrated in the improvement in accuracy associated with soil carbon stock estimation. Adopting a gridded approach to estimating D_b approximately halves the uncertainty associated with predictions. This alone should prompt further study into the potential of the technique to improve hydrological and agricultural models.

5.5 Conclusions

The two primary aims of this thesis were to predict soil D_b on a continuous grid and to introduce expert knowledge into the modelling process. The advantage of producing a gridded prediction of D_b is that data generated in this manner can be used to improve the prediction of models which require a spatial estimate of D_b as a parameter. This study has shown the utility of this output to landscape-scale carbon-stock estimation, which was predicted more accurately (particularly its systematic variability) using the

gridded method in comparison with a polygon-based average measure of D_b, which is the traditional method. Producing soil attribute data on a grid is useful, even if ultimately a soil class is required as the mapped output (such as in the ISIS project). This is because, provided the grid cell size is sufficiently small, the gridded prediction can be used to populate a soil classification system with attribute data. The major advantage of this approach it that sampling attribute data from a grid will provide both a mean value for each class and a confidence interval for the range of within-class values, thereby providing a measure of uncertainty in class-attribute data. The lack of such estimates has been a major criticism of the polygon-based soil mapping approach for decades. When a spatial representation of D_b is required, the use of a gridded model has been shown to be an improvement to the 'average-by-polygon' method which is currently used.

Work presented in this thesis has also shown that it is possible to use expert knowledge as a resource to produce a gridded estimate of D_b, however, the inclusion of expert knowledge in the modelling process must be carefully monitored, as it does not always yield improved predictions. By using a BN modelling approach, expert-knowledge was used to both provide the structure for data-mining techniques and used as a proxy for empirical data. Using expert knowledge to structure a BN, which was then used for data mining, did not improve model predictions. This was attributed to a failure of the model to adequately represent the complexity of the natural processes acting within the study area. However, when expert knowledge was used in a simplified BN, it was able to produce similar result to a data-mining method, identifying the same spatial trends in D_b, without the need for empirical data. The findings of this study suggest that expert knowledge is a valuable and viable resource for the spatial prediction of soil D_b (and

probably other soil properties too) although it can only be considered a preferable method when there is a lack of empirical data.

REFERENCES

Aalders, I., Hough, R. L. and Towers, W. (2011), Risk of erosion in peat soils - an investigation using Bayesian belief networks, *Soil Use and Management*, vol. 27, no. 4, pp. 538-549.

Adams, W. A. (1973), The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils, *Journal of Soil Science*, vol. 24, no. 1, pp. 10-17.

Aguilera, P. A., Fernandez, A., Fernandez, R., Rumi, R. and Salmeron, A. (2011), Bayesian networks in environmental modelling, *Environmental Modelling & Software*, vol. 26, no. 12.

Agyare, W. A., Park, S. J. and Vlek, P. L. G. (2007), Artificial neural network estimation of saturated hydraulic conductivity, *Vadose Zone Journal*, vol. 6, no. 2, pp. 423-431.

Aitkenhead, M. J. and Aalders, I. H. (2009), Predicting land cover using GIS, Bayesian and evolutionary algorithm methods, *Journal of environmental management*, vol. 90, no. 1, pp. 236-250.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. And Csaki, F. (Eds.), *International Symposium in Information Theory*, Budapest: Akademiai Kiado. pp. 267-281.

Amini, M., Abbaspour, K. C., Khademi, H., Fathianpour, N., Afyuni, M. and Schulin, R. (2005), Neural network models to predict cation exchange capacity in arid regions of Iran, *European Journal of Soil Science*, vol. 56, no. 4, pp. 551-559.

Arrouays, D., Bellamy, P.H. and Paustian, K. (2009), Soil Inventory and monitoring. Current issues and gaps, *European Journal of Soil Science*, vol. 60, no. 5, pp. 721-722.

Arya, L. M. and Paris, J. F. (1981), A Physicoempirical Model to Predict the Soil-Moisture Characteristic from Particle-Size Distribution and Bulk-Density Data, *Soil Science Society of America Journal*, vol. 45, no. 6, pp. 1023-1030.

Avery, B. W. (1980), Soil classification for England and Wales [higher categories], *Technical Monograph, Soil Survey of England and Wales*, no. 14.

Avery, B.W. and Bascomb, C.L. (1982), Soil Survey Laboratory Methods, *Soil Survey Technical Monograph, 6*, Rothamsted Experimental Station, Harpenden.

Batjes, N. H. (1996), Total carbon and nitrogen in the soils of the world, *European Journal of Soil Science*, vol. 47, no. 2, pp. 151-163.

Bayes, T. (1783), Essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London*, vol. 53, no. 1763, pp. 370–418.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E. D. and Goldschmitt, M. (2005), Digital soil mapping using artificial neural networks, *Journal of Plant Nutrition and Soil Science*, vol.168, no.1, 21-33.

Behrens, T. and Scholten, T. (2006a), Digital soil mapping in Germany - a review, *Journal of Plant Nutrition and Soil Science-Zeitschrift Fur Pflanzenernahrung Und Bodenkunde*, vol. 169, no. 3, pp. 434-443.

Behrens, T. and Scholten, T. (2006b), A comparison of data mining approaches in predictive soil mapping, In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), *Digital Soil Mapping*. Developments in Soil Science 31. Elsevier, Amsterdam pp. 658.

Bellamy, P., Loveland, P., Bradley, R., Lark, R. and Kirk, G. (2005), Carbon losses from all soils across England and Wales 1978-2003 RID A-4855-2011, *Nature*, vol. 437, no. 7056, pp. 245-248.

Ben-Dor, E. (2002), Quantitative remote sensing of soil properties, *Advances in Agronomy*, vol. 75, pp. 173-243.

Benites, V. M., Machado, P. L. O. A., Fidalgo, E. C. C., Coelho, M. R. and Madari, B. E. (2007), Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil, *Geoderma*, vol. 139, no. 1-2, pp. 90-97.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J.M., Hendry, M.A., Porson, A., Gedney, N., Mercado, L.M., Sitch, S., Blyth, E., Boucher, O., Cox, P.M., Grimmond, C.S.B and Harding, R. J. (2011), The Joint UK Land Environment Simulator (JULES), model description–Part 1: energy and water fluxes, *Geoscientific Model Development*, vol. 4, no. 3, pp. 677-699.

Beven, K. J., and Kirkby, M. J. (1979), A physically based, variable contributing area model of basin hydrology, *Hydrological Sciences Journal*, vol. 24, no. 1, pp. 43-69.

Bishop, C. M. (1995), Neural networks: a principled perspective, *Neural Networks - Producing Dependable Systems*, Birmingham: Aston University.

Bockheim, J. G., Gennadiyev, A. N., Hammer, R. D. and Tandarich, J. P. (2005), Historical development of key concepts in pedology, *Geoderma*, vol. 124, no. 1-2, pp. 23-36.

Böhner, J., Köthe, R., Conrad, O., Gross, J., Ringeler, A. and Selige, T. (2002) Soil regionalisation by means of terrain analysis and process parameterisation. In: Micheli, E., Nachtergaele, F., Montanarella, L. (Eds.), *Soil Classification 2001*. European Soil Bureau, Research Report No. 7, EUR 20398 EN, Luxembourg, pp. 213–222.

Böhner, J., and Selige, T. (2006). Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In Böhner, J., McCloy, K.R. and Strobl, J. (Eds.): *SAGA–Analyses and Modelling Applications.–Göttinger Geographische Abhandlungen*, 115, Göttingen, Germany: Verlag Goltze, pp. 13-28.

Borsuk, M.E. (2008) Ecological Informatics: Bayesian networks, In: Jorgensen, S.E. and Fath, B. (Eds.), *Encyclopaedia of Ecology*. Elsevier, Oxford, pp. 307-317.

Bossard, M., Feranec, J., and Otahel, J. (2000), *CORINE Land Cover Technical Guide — Addendum 2000*, European Environmental Agency Technical Report No. 40, Copenhagen, Denmark.

Braakhekke, M. C., Wutzler, T., Beer, C., Kattge, J., Schrumpf, M., Ahrens, B., Schöning, I., Hoosbeek, M.R., Kruijt, B., Kabat., P. and Reichstein, M (2013), Modeling the vertical soil organic matter profile using Bayesian parameter estimation, *Biogeosciences*, vol. 10, pp. 399-420.

Breiman, L. (2001), Random forests, *Machine Learning*, vol. 45, no. 1, pp. 5-32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and regression trees*, California:Wadsworth.

Bui, E.N. (2004), Soil Survey and a Knowledge System, *Geoderma*, vol. 120, pp. 17-26.

Bui, E. N., Henderson, B. L., and Viergever, K. (2006), Knowledge discovery from models of soil properties developed through data mining, *Ecological modelling*, vol. 191, no. 3, pp.431-446.

Bui, E. N., Lougldead, A., and Corner, R. (1999), Extracting soil-landscape rules from previous soil surveys, *Australian Journal of Soil Research*, vol. 37, no. 3, pp. 495-508.

Burgess, T., and Webster, R. (1980), Optimal interpolation and isarithmic mapping of soil properties, *Journal of Soil Science*, vol. 31, no. 2, pp. 333-341.

Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L. and Twardy, C. (2011a), Expert Status and Performance, *Plos One*, vol. 6, no. 7, pp. e22998.

Burgman, M., Carr, A., Godden, L., Gregory, R., McBride, M., Flander, L. and Maguire, L. (2011b), Redefining expertise and improving ecological judgment, *Conservation Letters*, vol. 4, no. 2, pp. 81-87.

Büttner, G., Feranec, F., and Jaffrain, G. (2002), *Corine land cover update 2000*, Technical report. Copenhagen: European Environment Agency.

Buttner, G., Steenmans, C., Bossard, M., Feranec, J. and Kolar, J. (2000), *Land Cover - Land use mapping within the European CORINE programme*, Dordrecht, Netherlands:Springer.

Cain, J. (2001), Planning improvements in natural resources management. In: *Guidelines for using Bayesian networks to support the planning and management of development programmes in the water sector and beyond*. Centre for Ecology and Hydrology, Wallingford, UK.

Calhoun, F. G., Smeck, N. E., Slater, B. L., Bigham, J. M. and Hall, G. F. (2001), Predicting bulk density of Ohio soils from morphology, genetic principles, and laboratory characterization data, *Soil Science Society of America Journal*, vol. 65, no. 3, pp. 811-819.

Carre, F., McBratney, A. B., Mayr, T. and Montanarella, L. (2007), Digital soil assessments: Beyond DSM, *Geoderma*, vol. 142, no. 1-2, pp. 69-79.

Cassman, K. G. (1999), Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. *Proceedings of the National Academy of Sciences*, vol. 96, no. 11, pp. 5952-5959.

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J. and Fealy, R. (2013), Are fine resolution digital elevation models always the best choice in digital soil mapping?, *Geoderma*, vol. 195, *195*, 111-121.

Charniak, E. (1991), Bayesian networks without tears. *AI magazine*, vol. 12, no. 4, pp. 50-63.

Chen, S. H. and Pollino, C. A. (2012), Good practice in Bayesian network modelling, *Environmental Modelling & Software*, vol. 37, pp. 134-145.

Chi, M. T. H. (2006), Two approaches to the study of experts' characteristics. In: K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance*, New York: Cambridge University Press, pp. 21-30.

Choy, S. L., O'Leary, R. and Mengersen, K. (2009), Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models, *Ecology*, vol. 90, no. 1, pp. 265-277.

Clarke, G.R. (1940), *Soil Survey of England and Wales: Field Handbook*, Oxford: University Press.

Clemen, R. T. and Winkler, R. L. (1985), Limits for the precision and value of information from dependent sources, *Operations Research*, vol. 33, no. 2, pp. 427-442.

Clemen, R. T. and Winkler, R. L. (1999), Combining probability distributions from experts in risk analysis. *Risk Analysis*, vol. 19, no. 2, pp. 187-203.

Cook, S. E., Corner, R. J., Groves, P. R., and Grealish, G. J. (1996), Use of airborne gamma radiometric data for soil mapping, *Soil Research*, vol. 34, no. 1, pp. 183-194.

Corner, R.J., Hickey, R.J. and Cook S.E. (2002), Knowledge based soil attribute mapping in GIS: the Expector method, *Transactions in GIS*, vol. 6, no. 4, pp. 383-402.

Correa, M., Bielza, C., and Pamies-Teixeira, J. (2009), Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process, *Expert systems with applications*, vol. 36, no. 3, pp. 7270-7279.

Coupé, V. M., and Van Der Gaag, L. C. (2002), Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, vol. 36, no. 4, pp. 323-356.

Creamer, R. E., Brennan, F., Fenton, O., Healy, M. G., Lalor, S. T. J., Lanigan, G. J., Regan, J.T. and Griffiths, B. S. (2010), Implications of the proposed Soil Framework Directive on agricultural systems in Atlantic Europe–a review, *Soil Use and Management*, vol. 26. No. 3, pp. 198-211.

Dale, M. B., McBratney, A. B. and Russell, J. S. (1989), On the role of expert systems and numerical taxonomy in soil classification, *Journal of Soil Science*, vol. 40, no. 2, pp. 223-234.

Daly, K. and Fealy, R. (2007), *Digital Soil Information System for Ireland Scoping Study (2005-S-DS-22-M1) Final Report*, Environmental Protection Agency: Ireland.

Das, K., and Vyas, O. P. (2010), A suitability study of discretization methods for associative classifiers, *International Journal of Computer Applications*, vol. 5, no. 10, pp. 46-51.

Dawson, J. J. C. and Smith, P. (2007), Carbon losses from soil and its consequences for land-use management, *Science of the Total Environment*, vol. 382, no. 2-3, pp. 165-190.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J. and Muys, B. (2005), Predictive quality of pedotransfer functions for estimating bulk density of forest soils, *Soil Science Society of America Journal*, vol. 69, no. 2, pp. 500-510.

Defra UK (2004), *The first soil action plan for England 2004-2006*, London: Department of the Environment, Food and Rural Affairs.

Degroot, M. H. (1988), A Bayesian View of Assessing Uncertainty and Comparing Expert Opinion, *Journal of Statistical Planning and Inference*, vol. 20, no. 3.

DeGruijter, J. J., Walvoort, D. J. J. and vanGaans, P. F. M. (1997), Continuous soil maps - A fuzzy set approach to bridge the gap between aggregation levels of process and distribution models, *Geoderma*, vol. 77, no. 2-4, pp. 169-195.

Delbecq, A. L., Van de Ven, A. H. and Gustafson, D. H. (1975), *Group techniques for program planning: A guide to nominal group and Delphi processes*, Glenview, IL: Scott, Foresman.

Dexter, A. R. (1988), Advances in Characterization of Soil Structure, *Soil & Tillage Research*, vol. 11, no. 3-4, pp. 199-238.

Diamond, J. and Sills, P. (2011), Soils of Co. Waterford, *Soil Survey Bulletin No. 44*, Carlow: Teagasc.

Dlamini, W. M. (2010), A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland, *Environmental Modelling & Software*, vol. 25, no. 2.

Dokuchaev, V.V. (1883), *The Russian Chernozem Report to the Free Economic Society (in Russian)*. Imperial Univ. of St. Petersburg, St. Petersburg, Russia.

Druzdzel, M. J. and van der Gaag, L. C. (2000), Building probabilistic networks: Where do the numbers come from? - Guest editors' introduction, *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 4, pp. 481-486.

Duda, R. O., and Hart, P. E. (1973), *Pattern classification and scene analysis (Vol. 3)*, New York: Wiley.

Elith, J., Leathwick, J.R. and Hastie, T. (2008), A Working Guide to Boosted Regression Trees, *Journal of Animal Ecology*, vol. 77, pp. 802-813.

Ellert, B. H. and Bettany, J. R. (1995), Calculation of organic matter and nutrients stored in soils under contrasting management regimes, *Canadian Journal of Soil Science*, vol. 75, no. 4, pp. 529-538.

Entz, T., and Chang, C. (1991), Evaluation of soil sampling schemes for geostatistical analyses: A case study for soil bulk density, *Canadian Journal of Soil Science*, vol. 71, no. 2, pp. 165-176.

ESRI (Environmental Systems Resource Institute), *ArcMap 9.3*. ESRI, Redlands, California, 2009.

Evans, E., Ramsbottom, D. M., Wicks, J. M., Packman, J.C. and Penning-Rowsell, E. C. (2002), Catchment Flood Management Plans and the Modelling and Decision Support Framework, *Proceedings of the Institution of Civil Engineers*, vol. 150, no. 1, pp. 43-48.

Faraway, J. J. (2002), Practical regression and ANOVA using R. Ann Arbor, MI, self published. http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

Farewell, T. S. and Farewell, D. M. (2010), Knowledge-based Soil Attribute Mapping in GIS: Corrections and Extensions to the Expector Method, *Transactions in Gis*, vol. 14, no. 2, pp. 183-192.

Fayyad, U. M. and Irani, K. B. (1993), Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, pp. 1022–1027.

Finke, P. A. (2012), On digital soil assessment with models and the Pedometrics agenda, *Geoderma*, vol. 171, pp. 3-15.

Fossitt, J. (2000), *A Guide to Habitats in Ireland*, Kilkenny: The Heritage Council.

Franklin, J. (1998), Predicting the distributions of shrub species in California chaparral and coastal sage communities from climate and terrain-derived variables, *Journal of Vegetation Science* vol. 9, pp. 733–48.

French, B. K. and Legg, B. J. (1979), Rothamsted irrigation 1964-76, *Journal of agricultural science (Cambridge)*, vol. 92, pp. 15-37.

Friedman, J.H. and Meulman, J.J. (2003), Multiple additive regression trees with application in epidemiology, *Statistics in Medicine*, vol. 22, no. 9, pp. 1365–1381.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997), Bayesian network classifiers, *Machine Learning*, vol. 29, no. 2-3.

Fuller, R. M., Smith, G. M., Sanderson, J. M., Hill, R. A. and Thomson, A. G. (2002), The UK Land Cover Map 2000: Construction of a parcel-based vector map from satellite images, *Cartographic Journal*, vol. 39, no. 1, pp. 15-25.

Gardiner, M.J. and Radford, T. (1980). *Ireland: General Soil Map. Second Edition. An Foras Talúntais (now Teagasc)*, Dublin, Ireland.

Gardiner, M. J. and Ryan, P. (1969). A new generalised soil map of Ireland and its land-use interpretation. *Irish Journal of Agricultural Research*, vol. 8, pp. 95-109.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005), Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association,* vol. 100, no. 470, pp. 680-700.

Gigerenzer, G. and Hoffrage, U. (1995), How to improve Bayesian reasoning without instruction: Frequency formats, *Psychological review,* vol. 102, no. 4, pp. 684-704.

Goidts, E., van Wesemael, B. and Crucifix, M. (2009), Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales, *European Journal of Soil Science,* vol. 60, no. 5.

Goodale, C. L., Aber, J. D. And Ollinger, S. V. (1998), Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model, *Climate Research,* vol. 10, pp. 35-49.

Goovaerts, P. (1999), Geostatistics in soil science: state-of-the-art and perspectives, *Geoderma,* vol. 89, no. 1-2, pp. 1-45.

Gret-Regamey, A. and Straub, D. (2006), Spatially explicit avalanche risk assessment linking Bayesian networks to a GIS, *Natural Hazards and Earth System Sciences,* vol. 6, no. 6, pp. 911-926.

Grimm, R., Behrens, T., Maerker, M. and Elsenbeer, H. (2008), Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis, *Geoderma,* vol. 146, no. 1-2, pp. 102-113.

Grunwald, S. (2009), Multi-criteria characterization of recent digital soil mapping and modeling approaches, *Geoderma,* vol. 152, no. 3-4, pp. 195-207.

Grunwald, S., Thompson, J. A. and Boettinger, J. L. (2011), Digital soil mapping and modeling at continental scales: Finding solutions for global issues, *Soil Science Society of America Journal*, vol. 75, no. 4, pp. 1201-1213.

Haering, T., Dietz, E., Osenstetter, S., Koschitzki, T. and Schroeder, B. (2012), Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils, *Geoderma,* vol. 185, pp. 37-47.

Hallett, S.H., Hollis, J.M., and Keay, C.A. (1998) Derivation and evaluation of a set of pedogenically-based empirical algorithms for predicting bulk density in British soils. Research paper available from http://www.landis.org.uk/downloads/index.cfm_Predicting_Bulk _Density.pdf

Hallett, S. H. and Jones, R. J. A. (1993), Compilation of an Accumulated Temperature Database for use in an Environmental Information-System, *Agricultural and Forest Meteorology,* vol. 63, no. 1-2, pp. 21-34.

Hanegraaf, M. C., Hoffland, E., Kuikman, P. J. and Brussaard, L. (2009), Trends in soil organic matter contents in Dutch grasslands and maize fields on sandy soils, *European Journal of Soil Science,* vol. 60, no. 2, pp. 213-222.

Hansen, M. K., Brown, D. J., Dennison, P. E., Graves, S. A. and Bricklemyer, R. S. (2009), Inductively mapping expert-derived soil-landscape units within dambo wetland catenae using multispectral and topographic data, *Geoderma*, vol. 150, no. 1, pp. 72-84.

Harrison, R. G., Jones, C. D. and Hughes, J. K. (2008), Competing roles of rising CO2 and climate change in the contemporary European carbon balance, *Biogeosciences*, vol. 5, no. 1, pp. 1-10.

Hartemink, A. E., Lowery, B. and Wacker, C. (2012), Soil maps of Wisconsin, *Geoderma*, vol. 189, pp. 451-461.

Hartemink, A. E. and McBratney, A. (2008), A soil science renaissance, *Geoderma*, vol. 148, no. 2, pp. 123-129.

Heckerman, D. (1997), Bayesian networks for data mining, *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 79-119.

Henderson, B. L., Bui, E. N., Moran, C. J. and Simon, D. A. P. (2005). Australia-wide predictions of soil properties using decision trees. *Geoderma*, vol. 124, no. 3, pp. 383-398.

Hess, T. M. (2000) *Reference Evapotranspiration Program*, Cranfield University, Silsoe.

Heuscher, S. A., Brandt, C. C. and Jardine, P. M. (2005), Using soil physical and chemical properties to estimate bulk density, *Soil Science Society of America Journal*, vol. 69, no. 1, pp. 51-56.

Heuvelink, G. B. M. and Webster, R. (2001), Modelling soil variation: past, present, and future, *Geoderma*, vol. 100, no. 3-4, pp. 269-301.

Hodgkinson, G. P., Bown, N. J., Maule, A. J., Glaister, K. W. and Pearman, A. D. (1999), Breaking the frame: An analysis of strategic cognition and decision making under uncertainty, *Strategic Management Journal*, vol. 20, no. 10, pp. 977-985.

Hodgson, J.M. (1976), Soil survey field handbook. Soil Survey England and Wales; Harpenden, *Technical Monograph No. 5*.

Hollis, J. M., Hannam, J. and Bellamy, P. H. (2012), Empirically-derived pedotransfer functions for predicting bulk density in European soils, *European Journal of Soil Science*, vol. 63, no. 1, pp. 96-109.

Holmberg, M., Forsius, M., Starr, M., and Huttunen, M. (2006), An application of artificial neural networks to carbon, nitrogen and phosphorus concentrations in three boreal streams and impacts of climate change, *Ecological modelling*, vol. 195, no 1, pp. 51-60.

Hough, R. L., Towers, W. and Aalders, I. (2010), The Risk of Peat Erosion from Climate Change: Land Management CombinationsAn Assessment with Bayesian Belief Networks, *Human and Ecological Risk Assessment*, vol. 16, no. 5, pp. 962-976.

Hudson, B. D. (1992), The Soil Survey as Paradigm-Based Science, *Soil Science Society of America Journal*, vol. 56, no. 3, pp. 836-841.

INSPIRE EU Directive, (2007), Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in

the European Community (INSPIRE). *Official Journal of the European Union*, L 108/1 50. http://inspire.jrc.ec.europa.eu/

IUSS Working Group WRB (2006), World Reference Base for Soil Resources, second ed. *World Soil Resources Report 103*. FAO, Rome.

Iwahashi, J. and Pike, R. J. (2007), Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature, *Geomorphology,* vol. 86, no. 3-4, pp. 409-440.

Jakeman, A. J., Letcher, R. A. and Norton, J. P. (2006), Ten iterative steps in development and evaluation of environmental models, *Environmental Modelling & Software,* vol. 21, no. 5, pp. 602-614.

Jalabert, S. S. M., Martin, M. P., Renaud, J. -., Boulonne, L., Jolivet, C., Montanarella, L. and Arrouays, D. (2010), Estimating forest soil bulk density using boosted regression modelling, *Soil Use and Management,* vol. 26, no. 4, pp. 516-528.

Janssens, I. A., Freibauer, A., Schlamadinger, B., Ceulemans, R., Ciais, P., Dolman, A. J., Heimann, M., Nabuurs, G. J., Smith, P., Valentini, R. and Schulze, E. D. (2005), The carbon budget of terrestrial ecosystems at country-scale - a European case study, *Biogeosciences,* vol. 2, no. 1, pp. 15-26.

Jenny, H. (1941) *Factors of soil formation*. McGraw-Hill: New York, USA.

Jensen, F. V. (1996), *An Introduction to Bayesian Networks*, London: UCL Press.

Jensen, F.V. (2001) Bayesian Networks and Decision Graphs, New York: Springer-Verlag.

Jensen, F.V. and Andersen, S.K. (1990), Approximations in Bayesian belief universes for knowledge-based systems. In: *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pp. 162-169.

Jiang, L., Zhang, H., Cai, Z. and Su, J. (2005). Learning tree augmented naive Bayes for ranking. In *Proceedings of the 10th International Conference on Database Systems for Advanced Applications*, Berlin: Springer-Verlag, pp. 688–698.

Johnson, S., Low-Choy, S. and Mengersen, K. (2012), Integrating Bayesian networks and geographic information systems: good practice examples., *Integrated environmental assessment and management,* vol. 8, no. 3.

Jones, R. J. A., Hiederer, R., Rusco, E. and Montanarella, L. (2005), Estimating organic carbon in the soils of Europe for policy support, *European Journal of Soil Science,* vol. 56, no. 5, pp. 655-671.

Jones, R. J. A., and Thomasson, A.J. (1985) *An Agroclimatic Databank for England and Wales*, Technical Monograph,16, Soil Survey, Harpenden.

Kadane, J. B. and Wolfson, L. J. (1998), Experiences in elicitation, *Journal of the Royal Statistical Society Series D-the Statistician,* vol. 47, no. 1, pp. 3-19.

Kallis, G. and Butler, D. (2001), The EU water framework directive: measures and implications, *Water policy*, vol. 3, no. 2, pp. 125-142.

Katterer, T., Andren, O. and Jansson, P. -. (2006), Pedotransfer functions for estimating plant available water and bulk density in Swedish agricultural soils, *Acta Agriculturae Scandinavica Section B-Soil and Plant Science*, vol. 56, no. 4, pp. 263-276.

Kaur, R., Kumar, S. and Gurung, H. P. (2002), A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs, *Australian Journal of Soil Research*, vol. 40, no. 5, pp. 847-857.

Keshavarzi, A., Sarmadian, F., Sadeghnejad, M. and Pezeshki, P. (2010), Developing pedotransfer functions for estimating some soil properties using artificial neural network and multivariate regression approaches, *ProEnvironment/ProMediu*, vol. 3, no. 6, pp. 322-330.

King, A.W., Dilling, L., Zimmerman, G.P., Fairman, D.M., Houghton, R.A., Marland, G., Rose, A.Z., Wilbanks, J. (2007), The First State of the Carbon Cycle Report (SOCCR): The North American Carbon Budget and Implications for the Global Carbon Cycle, A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. National Oceanic and Atmospheric Administration, National Climatic Data Center, Asheville, NC, USA, pp. 1-14.

Kline, J.R., (1973), Mathematical simulation of soil–plant relationships and soil genesis, *Soil Science*, vol. 115, pp. 240–249.

Knol, A. B., Slottje, P., van der Sluijs, J. P. and Lebret, E. (2010), The use of expert elicitation in environmental health impact assessment: a seven step procedure, *Environmental Health*, vol. 9, pp. 19.

Krueger, T., Page, T., Hubacek, K., Smith, L. and Hiscock, K. (2012), The role of expert opinion in environmental modelling, *Environmental Modelling & Software*, vol. 36.

Kuhn, M. (2008), Building predictive models in R using the caret package, *Journal of Statistical Software*, vol. 28, no. 5, pp. 1-26.

Kuhnert, P. M. and Hayes, K. R. (2009), How believable is your BBN? In *Proceedings of the 18th World IMACS/MODSIM Congress, Cairns, Australia*, pp. 13-17.

Kuhnert, P. M., Martin, T. G. and Griffiths, S. P. (2010), A guide to eliciting and using expert knowledge in Bayesian ecological models, *Ecology Letters*, vol. 13, no. 7, pp. 900-914.

Kynn, M. (2008), The 'heuristics and biases' bias in expert elicitation, *Journal of the Royal Statistical Society Series A-Statistics in Society*, vol. 171, pp. 239-264.

Lampurlanes, J. and Cantero-Martinez, C. (2003), Soil bulk density and penetration resistance under different tillage and crop management systems and their relationship with barley root growth, *Agronomy Journal*, vol. 95, no. 3, pp. 526-536.

Lark, R. M., Bishop, T. F. A. and Webster, R. (2007), Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties, *Geoderma*, vol. 138, no. 1, pp. 65-78.

Lawley, R. and Smith, B. (2008), Digital soil mapping at a national scale: a knowledge and GIS based approach to improving parent material and property information. In: Hartemink, A.E., McBratney, A.B., Mendonça-Santos, M.L. (Eds.), *Digital Soil Mapping with Limited Data*. Springer, Dordrecht, pp. 173–182.

Lee, J., Hopmans, J.W., Rolston, D.E., Baer, S.G. and Six, J. (2009), Determining Carbon Stock Changes: Simple Bulk Density Corrections Fail, *Agriculture, Ecosystems and Environment*, vol. 134, no. 3, pp. 251-256.

Lee, M., Choi, J., Oh, H., Won, J., Park, I. and Lee, S. (2012), Ensemble-based landslide susceptibility maps in Jinbu area, Korea, *Environmental Earth Sciences*, vol. 67, no. 1, pp. 23-37.

Lemercier, B., Lacoste, M., Loum, M. and Walter, C. (2012), Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach, *Geoderma*, vol. 171.

Liaw, A., and Wiener, M. (2002) Classification and Regression by randomForest. R News: *The Newsletter of the R Project* (http://cran.r-project.org/doc/Rnews/), vol. 2, no. 3, pp. 18–22.

Ließ, M., Glaser, B. and Huwe, B. (2012), Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models, *Geoderma*, vol. 170, no. 0, pp. 70-79.

Liu, H., Hussain, F., Tan, C. L. and Dash, M. (2002), Discretization: An enabling technique, *Data Mining and Knowledge Discovery*, vol. 6, no. 4, pp. 393-423.

Liu, T. L., Juang, K. W. and Lee, D. Y. (2006), Interpolating soil properties using kriging combined with categorical information of soil maps, *Soil Science Society of America Journal*, vol. 70, no. 4, pp. 1200-1209.

Lou, W. G. and Nakai, S. (2001), Application of artificial neural networks for predicting the thermal inactivation of bacteria: a combined effect of temperature, pH and water activity, *Food Research International*, vol. 34, no. 7, pp. 573-579.

Loveland, P. J. (1990), *The National Soil Inventory of England and Wales UK*.

Low Choy, S., O'Leary, R. and Mengersen, K. (2009), Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models, *Ecology*, vol. 90, no. 1, pp. 265-277.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, vol. 10, no. 4, pp. 325-337.

Mackney D., Hodgson J.M., Hollis J.M. and Staines S.J. (1983), Legend for the 1 : 250,000 Soil Map of England and Wales. *Soil Survey of England and Wales*: Harpenden, pp.21.

Marcot, B. G., Holthausen, R. S., Raphael, M. G., Rowland, M. M., and Wisdom, M. J. (2001), Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest ecology and management*, vol. 153, no. 1, pp. 29-42.

Marcot, B. G., Steventon, J. D., Sutherland, G. D. and McCann, R. K. (2006), Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation, *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere*, vol. 36, no. 12, pp. 3063-3074.

Martin, M. P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K. M., Bourgeon, G. and Arrouays, D. (2009), Optimizing Pedotransfer Functions for Estimating Soil Bulk Density Using Boosted Regression Trees, *Soil Science Society of America Journal*, vol. 73, no. 2, pp. 485-493.

Martin, M. P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L. and Arrouays, D. (2011), Spatial distribution of soil organic carbon stocks in France, *Biogeosciences*, vol. 8, no. 5, pp. 1053-1065.

Martin, T. G., Burgman, M. A., Fidler, F., Kuhnert, P. M., Low-Choy, S., McBride, M. and Mengersen, K. (2012), Eliciting expert knowledge in conservation science, *Conservation Biology*, vol. 26, no. 1, pp. 29-38.

Matthews, E., Payne, R., Rohweder, M. and Murray, S. (2000), *Pilot Analysis of Global Ecosystems (PAGE): Forest Ecosystems*, Washington: World Resources Institute.

Mayr, T. and Palmer, R.C. (2007), Digital soil mapping: an England and Wales perspective, In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), *Digital Soil Mapping. An Introductory Perspective*. Developments in Soil Science, vol. 31. Elseveier, Amsterdam, pp. 365–375.

Mayr, T.R., Palmer, R.C. and Cooke, H.J. (2008) Digital Soil Mapping Using Legacy Data in the Eden Valley, UK, In: Hartemink, A.E., McBratney, A.B., de Lourdes Mendonça Santos, M. (Eds.), *Digital Soil Mapping With Limited Data*. Springer, Netherlands, pp. 291-301.

Mayr, T., Rivas-Casado, M., Bellamy, P., Palmer, R., Zawadzka, J. and Corstanje, R. (2010), Two methods for using legacy data in digital soil mapping, In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*, Springer, Dordrecht, pp. 191–202.

McBratney, A.B., Minasny, B., Cattle, S.R. and Vervoort, R.W. (2002), From Pedotransfer Functions to Soil Inference Systems. *Geoderma* vol. 109, pp. 41-73.

McBratney, A. B. and Odeh, I. O. A. (1997), Application of fuzzy sets in soil science: Fuzzy logic, fuzzy measurements and fuzzy decisions, *Geoderma*, vol. 77, no. 2-4.

McBratney, A. B., Santos, M. L. M. and Minasny, B. (2003), On digital soil mapping, *Geoderma*, vol. 117, no. 1-2, pp. 3-52.

McBratney, A. B., Webster, R. and Burgess, T. M. (1981), The design of optimal sampling schemes for local estimation and mapping of of regionalized variables- I: Theory and method, *Computers & Geosciences*, vol. 7, no. 4, pp. 331-334.

McBride, M. F., and Burgman, M. A. (2012), What is expert knowledge, how is such knowledge gathered, and how do we use it to address questions in landscape ecology? In Perera, A., Johnson, C. and Drew, C. A. (Eds.), *Expert knowledge and its application in landscape ecology*. New York: Springer-Verlag.

McCann, R. K., Marcot, B. G. and Ellis, R. (2006), Bayesian belief networks: applications in ecology and natural resource management, *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere*, vol. 36, no. 12, pp. 3053-3062.

McCoy, J. and Johnston, K. (2002), *Using ArcGIS spatial analyst*, Redlands: ESRI.

McCloskey, J. T., Lilieholm, R. J. and Cronan, C. (2011), Using Bayesian belief networks to identify potential compatibilities and conflicts between development and landscape conservation, *Landscape and Urban Planning*, vol. 101, no. 2, pp. 190-203.

McGrath, S.P. and Loveland, P.J. (1992), *The Soil Geochemical Atlas of England and Wales*. Blackie, Glasgow.

McKenzie, N. J. and Ryan, P. J. (1999), Spatial prediction of soil properties using environmental correlation. *Geoderma*, vol. 89, no. 1, pp. 67-94.

Mestdagh, I., Sleutel, S., Lootens, P., Van Cleemput, O., Beheydt, D., Boeckx, P., De Neve, S., Hofman, G., Van Camp, N., Vande Walle, I., Samson, R., Verheyen, K., Lemeur, R. and Carlier, L. (2009), Soil organic carbon-stock changes in Flemish grassland soils from 1990 to 2000, *Journal of Plant Nutrition and Soil Science-Zeitschrift Fur Pflanzenernahrung Und Bodenkunde*, vol. 172, no. 1, pp. 24-31.

Miller, D. A. and White, R. A. (1998), A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modelling, *Earth Interactions*, vol. 2, no. 2, pp. 1-26.

Minasny, B., McBratney, A. B. and Bristow, K. L. (1999), Comparison of different approaches to the development of pedotransfer functions for water-retention curves, *Geoderma*, vol. 93, no. 3-4, pp. 225-253.

Minasny, B., McBratney, A. B., Tranter, G. and Murphy, B. W. (2008), Using soil knowledge for the evaluation of mid-infrared diffuse reflectance spectroscopy for predicting soil physical and mechanical properties, *European Journal of Soil Science*, vol. 59, no. 5, pp. 960-971.

Minasny, B. and Hartemink, A. E. (2011), Predicting soil properties in the tropics, *Earth-Science Reviews*, vol. 106, no. 1-2, pp. 52-62.

Moore, I. D. and Burch, G. J. (1986), Sediment Transport Capacity of Sheet and Rill Flow - Application of Unit Stream Power Theory, *Water Resources Research*, vol. 22, no. 8, pp. 1350-1360.

Mora-Vallejo, A., Claessens, L., Stoorvogel, J. and Heuvelink, G. B. M. (2008), Small scale digital soil mapping in Southeastern Kenya, *Catena*, vol. 76, no. 1, pp. 44-53.

Moreira, C. S., Brunet, D., Verneyre, L., Sa, S. M. O., Galdos, M. V., Cerri, C. C. and Bernoux, M. (2009), Near infrared spectroscopy for soil bulk density assessment, *European Journal of Soil Science*, vol. 60, no. 5, pp. 785-791.

Murray, J. V., Stokes, K. E. and van Klinken, R. D. (2012), Predicting the potential distribution of a riparian invasive plant: the effects of changing climate, flood regimes and land-use patterns, *Global Change Biology*, vol. 18, no. 5, pp. 1738-1753.

Nachtergaele, F.O. and Van Ranst, E. (2003), Qualitative and quantitative aspects of soil databases in tropical countries. In: Stoops, G. (Ed.), *Evolution of Tropical Soil Science: Past and Future*, Brussel: Koninklijke Academie voor Overzeese Wetenschappen, pp. 107–126.

Nadkarni, S. and Shenoy, P. P. (2004), A causal mapping approach to constructing Bayesian networks, *Decision Support Systems*, vol. 38, no. 2, pp. 259-281.

Naylor, D. (1978), The geological survey of Ireland, *Irish Geography*, vol. 11, no. 1, pp. 155-160.

Norsys Software Corp (2012) Netica 5.05. Available at: http://www.norsys.com

Norton, L., Elliott, J. A., Maberly, S. C. And May, L. (2012), Using models to bridge the gap between land use and algal blooms: An example from the Loweswater catchment, UK. *Environmental Modelling & Software*, vol. 36, pp. 64-75.

O'Hagan, A. and Oakley, J. E. (2004), Probability is perfect, but we can't elicit it perfectly, *Reliability Engineering & System Safety*, vol. 85, no. 1-3, pp. 239-248.

O'Hagan, A. (2012), Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux, *Environmental Modelling & Software*, vol. 36, pp. 35-48.

O'Leary, R. A., Low Choy, S. J., Murray, J. V., Kynn, M., Denham, R., Martin, T. G. and Mengersen, K. (2009) Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby *petrogale penicillata*. *Environmetrics*, vol. 20, no. 4, pp. 379–398.

Olaya, V. and Conrad, O. (2009), Geomorphometry in SAGA. In: Hengl, T., Reuter, H. (Eds.), *Geomorphometry - Concepts, Software, Applications*. *Vol. 33 of Developments in Soil Science*. Amsterdam: Elsevier, pp. 293–308.

Pachepsky, Y. A., Guber, A., Jacques, D., Simunek, J., Van Genuchten, M. T., Nicholson, T. and Cady, R. (2006), Information content and complexity of simulated soil water fluxes, *Geoderma*, vol. 134, no. 3-4, pp. 253-266.

Pachepsky, Y. A., Timlin, D. and Varallyay, G. (1996), Artificial neural networks to estimate soil water retention from easily measurable data, *Soil Science Society of America Journal*, vol. 60, no. 3, pp. 727-733.

Park, S. J., Hwang, C. S. and Vlek, P. L. G. (2005), Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions, *Agricultural Systems*, vol. 85, no. 1, pp. 59-81.

Pearl, J. (1988), *Probabilistic reasoning in intelligent systems: networks of plausible inference*, San Francisc:o Morgan Kaufmann.

Pennock, D. J. and Veldkamp, A. (2006), Advances in landscape-scale soil research, *Geoderma*, vol. 133, no. 1, pp. 1-5.

Pennock, D. J., Zebarth, B. J. and Dejong, E. (1987), Landform Classification and Soil Distribution in Hummocky Terrain, Saskatchewan, Canada, *Geoderma,* vol. 40, no. 3-4, pp. 297-315.

Perry, M. and Hollis, D. (2005), The generation of monthly gridded datasets for a range of climatic variables over the UK, *International Journal of Climatology,* vol. 25, no. 8, pp. 1041-1054.

Pires, L.F., Rosa, J.A., Pereira, A.B., Arthur, R.C.J. and Bacchi, O.O.S. (2009), Gamma-ray attenuation method as an efficient tool to investigate soil bulk density spatial variation, *Annals of Nuclear Energy*, vol. 36, no. 11, pp. 1734-1739.

Prasad, A. M., Iverson, L. R. and Liaw, A. (2006), Newer classification and regression tree techniques: Bagging and random forests for ecological prediction, *Ecosystems,* vol. 9, no. 2, pp. 181-199.

Preston, R. and Mills, P. (2002), Generation of a Hydrologically Corrected Digital Elevation Model for the Republic of Ireland. *Unpublished report submitted to EPA by Compass Informatics as part of the 2000-LS-2.2, 2*.

Qi, F., Zhu, A., Harrower, M. and Burt, J. E. (2006), Fuzzy soil mapping based on prototype category theory, *Geoderma,* vol. 136, no. 3, pp. 774-787.

Rawlins, B. G., Marchant, B. P., Smyth, D., Scheib, C., Lark, R. M. and Jordan, C. (2009), Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland, *European Journal of Soil Science,* vol. 60, no. 1, pp. 44-54.

Rawls, W. J. (1983), Estimating Soil Bulk-Density from Particle-Size Analysis and Organic-Matter Content, *Soil Science,* vol. 135, no. 2, pp. 123-125.

Razakamanarivo, R. H., Grinand, C., Razafindrakoto, M. A., Bernoux, M. and Albrecht, A. (2011), Mapping organic carbon stocks in eucalyptus plantations of the central highlands of Madagascar: A multiple regression approach. *Geoderma,* vol. 162, no. 3, pp. 335-346.

Renooij, S. (2001), Probability elicitation for belief networks: issues to consider, *Knowledge Engineering Review,* vol. 16, no. 3, pp. 255-269.

Robinson, J. W. and Hartemink, A. J. (2010), Learning non-stationary dynamic Bayesian networks. *The Journal of Machine Learning Research*, vol. 11, pp. 3647-3680.

Rossel, R. A. V. and Behrens, T. (2010), Using data mining to model and interpret soil diffuse reflectance spectra, *Geoderma,* vol. 158, no. 1-2.

Rossiter, D. (2005), Digital Soil Mapping: Towards a Multiple Use Soil Information System. *Análisis Geográficos (Revista del Instituto Geográfico "Augusín Codazzi")*, vol. 32, no.1, pp. 7–15.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986), Learning Representations by Back-Propagating Errors, *Nature*, vol. 323, no. 6088, pp. 533-536.

Sanchez, P. A., Ahamed, S., Carre, F., Hartemink, A. E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A. B., McKenzie, N. J., Mendonca-Santos, M. d. L., Minasny, B., Montanarella, L., Okoth, P., Palm, C. A., Sachs, J. D., Shepherd, K. D., Vagen, T., Vanlauwe, B., Walsh, M. G., Winowiecki, L. A. and Zhang, G. (2009), Digital Soil Map of the World, *Science*, vol. 325, no. 5941, pp. 680-681.

Schrumpf, M., Schulze, E. D., Kaiser, K. and Schumacher, J. (2011), How accurately can soil organic carbon stocks and stock changes be quantified by soil inventories?, *Biogeosciences*, vol. 8, no. 5, pp. 1193-1212.

Scull, P., Franklin, J., and Chadwick, O.A. (2005), The Application of Classification Tree Analysis to Soil Type Prediction in a Desert Landscape, *Ecological Modelling*, vol. 181, no. 1, pp. 1-15.

Scull, P., Franklin, J., Chadwick, O. A. and McArthur, D. (2003), Predictive soil mapping: a review, *Progress in Physical Geography*, vol. 27, no. 2, pp. 171-197.

Shi, X., Long, R., Dekett, R. and Philippe, J. (2009), Integrating Different Types of Knowledge for Digital Soil Mapping, *Soil Science Society of America Journal*, vol. 73, no. 5, pp. 1682-1692.

Skidmore, A.K., Watford, F., Luckananurug, P. and Ryan, P.J. (1996), An Operational GIS Expert System for Mapping Forest Soils, *Photogrammetric Engineering & Remote Sensing*, vol. 62, no. 5, pp. 501-511.

Smith, C. S., Howes, A. L., Price, B. and McAlpine, C. A. (2007), Using a Bayesian belief network to predict suitable habitat of an endangered mammal - The Julia Creek dunnart (Sminthopsis douglasi), *Biological Conservation*, vol. 139, no. 3-4, pp. 333-347.

Smith, P., Andren, O., Karlsson, T., Perala, P., Regina, K., Rounsevell, M. and van Wesemael, B. (2005), Carbon sequestration potential in European croplands has been overestimated, *Global Change Biology*, vol. 11, no. 12, pp. 2153-2163.

Smith, P., Milne, R., Powlson, D. S., Smith, J. U., Falloon, P. and Coleman, K. (2006), Revised estimates of the carbon mitigation potential of UK agricultural land, *Soil Use and Management*, vol. 16, no. 4, pp. 293-295.

Soil survey of England and Wales, (1983), *Soil Map of England and Wales, Scale 1:250000*. Harpenden, UK.

Spiegelhalter, D. J., Franklin, R. C. and Bull, K. (1990), Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system. In *Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence*, North-Holland Publishing Co, pp. 285-294.

StatSoft, Inc. (2011), *Electronic Statistics Textbook*, Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/.

Steller, R.M., Jelinski, N.A. and Kucharik, C.J. (2008), Developing Models to Predict Soil Bulk Density in Southern Wisconsin using Soil Chemical Properties, *Journal of Integrative Biosciences*, vol. 6, no. 1, pp. 53-63.

Stoorvogel, J. J., Kempen, B., Heuvelink, G. B. M. and de Bruin, S. (2009), Implementation and evaluation of existing knowledge for digital soil mapping in Senegal, *Geoderma*, vol. 149, no. 1-2, pp. 161-170.

Suuster, E., Ritz, C., Roostalu, H., Kolli, R. and Astover, A. (2012), Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data, *European Journal of Soil Science*, vol. 63, no. 3, pp. 351-359.

Taalab, K. P., Corstanje, R., Creamer, R. and Whelan, M. J. (2012), Modeling soil bulk density at the landscape scale and its contributions to C stock uncertainty, *Biogeosciences Discuss*, vol. 9, pp. 18831-18864.

Tavares Wahren, F., Tarasiuk, M., Mykhnovych, A., Kit, M., Feger, K. H. and Schwärzel, K. (2012), Estimation of spatially distributed soil information: dealing with data shortages in the Western Bug Basin, Ukraine, *Environmental Earth Sciences*, vol. 65. No. 5, pp. 1501-1510.

Throop, H. L., Archer, S. R., Monger, H. C. and Waltman, S. (2012), When bulk density methods matter: Implications for estimating soil organic carbon pools in rocky soils, *Journal of Arid Environments*, vol. 77, pp. 66-71.

Tornquist, C. G., Giasson, E., Mielniczuk, J., Pellegrino Cerri, C. E. and Bernoux, M. (2009), Soil Organic Carbon Stocks of Rio Grande do Sul, Brazil RID B-3090-2008, *Soil Science Society of America Journal*, vol. 73, no. 3, pp. 975-982.

Tranter, G., Minasny, B., Mcbratney, A. B., Murphy, B., Mckenzie, N. J., Grundy, M. and Brough, D. (2007), Building and testing conceptual and empirical models for predicting soil bulk density, *Soil Use and Management*, vol. 23, no. 4, pp. 437-443.

Tranter, G., Minasny, B., Mcbratney, A. B., Murphy, B., Mckenzie, N. J., Grundy, M. and Brough, D. (2007), Building and testing conceptual and empirical models for predicting soil bulk density, *Soil Use and Management*, vol. 23, no. 4, pp. 437-443.

Ungaro, F., Staffilani, F. and Tarocco, P. (2010), Assessing and Mapping Topsoil Organic Carbon Stock at Regional Scale: a Scorpan Kriging Approach Conditional on Soil Map Delineations and Land use, *Land Degradation & Development*, vol. 21, no. 6, pp. 565-581.

Utset, A., Lopez, T. and Diaz, M. (2000), A comparison of soil maps, kriging and a combined method for spatially predicting bulk density and field capacity of ferralsols in the Havana-Matanzas Plain, *Geoderma*, vol. 96, no. 3, pp. 199-213.

Van der Gaag, L. C., Renooij, S., Witteman, C. L., Aleman, B. M. And Taal, B. G. (1999), How to elicit many probabilities. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., pp. 647-654.

Veronesi, F., Corstanje, R. and Mayr, T. (2012), Mapping soil compaction in 3D with depth functions, *Soil & Tillage Research,* vol. 124, pp. 111-118.

Voltz, M. and Webster, R. (1990), A Comparison of Kriging, Cubic-Splines and Classification for Predicting Soil Properties from Sample Information, *Journal of Soil Science,* vol. 41, no. 3, pp. 473-490.

Walter, C., Lagacherie and P., Follain, S. (2006) Integrating pedological knowledge into soil digital mapping. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), *Digital Soil Mapping: An introductory perspective*. Elsevier, Amsterdam, pp. 281–300.

Webster, R. (2000), Is soil variation random?, *Geoderma*, vol. 97, no.3, pp. 149-163.

Webster, R., Harrod, T.R., Staines, S.J. and Hogan, D.V. (1979), Grid Sampling and Computer Mapping of the Ivybridge Area, Devon. *Soil Survey Technical Monograph, 12*, Harpenden.

Wiesmeier, M., Barthold, F., Blank, B. and Koegel-Knabner, I. (2011), Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem, *Plant and Soil,* vol. 340, no. 1-2, pp. 7-24.

Wosten, J. H. M., Pachepsky, Y. A. and Rawls, W. J. (2001), Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics, *Journal of Hydrology,* vol. 251, no. 3-4, pp. 123-150.

Wutzler, T., Wirth, C., and Schumacher, J. (2008), Generic biomass functions for Common beech (Fagus sylvatica) in Central Europe: predictions and components of uncertainty, *Canadian Journal of Forest Research,* vol. 38, no. 6, pp. 1661-1675.

Yamada, K., Elith, J., McCarthy, M. and Zerger, A. (2003), Eliciting and integrating expert knowledge for wildlife habitat modelling, *Ecological Modelling,* vol. 165, no. 2-3, pp. 251-264.

Yu, J., Wang, Y., Li, Y., Dong, H., Zhou, D., Han, G., Wu, H., Wang, G., Mao, P. and Gao, Y. (2012), Soil organic carbon storage changes in coastal wetlands of the modern Yellow River Delta from 2000 to 2009, *Biogeosciences,* vol. 9, no. 6.

Zaehle, S., Bondeau, A., Carter, T. R., Cramer, W., Erhard, M., Prentice, I. C., Reginster, I., Rounsevell, M. D. A., Sitch, S., Smith, B., Smith, P. C. and Sykes, M. (2007), Projected changes in terrestrial carbon storage in Europe under climate and land-use change, 1990-2100, *Ecosystems,* vol. 10, no. 3, pp. 380-401.

Zhao, Y.-C., Shi, X.-Z. (2010) Spatial prediction and uncertainty assessment of soil organic carbon in Hebei province, China. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*, Springer, Dordrecht, pp. 227–239.

Zhao, Z., Yang, Q., Benoy, G., Chow, T. L., Xing, Z., Rees, H. W. and Meng, F. (2010), Using artificial neural network models to produce soil organic carbon content distribution maps across landscapes, *Canadian Journal of Soil Science,* vol. 90, no. 1, pp. 75-87.

Zhu, A. X. (2000), Mapping soil landscape as spatial continua: the neural network approach, *Water Resources Research*, vol. 36, no. 3, pp. 663-677.

Zhu, A. X., Hudson, B., Burt, J., Lubich, K. and Simonson, D. (2001), Soil mapping using GIS, expert knowledge, and fuzzy logic, *Soil Science Society of America Journal*, vol. 65, no. 5, pp. 1463-1472.

Ziadat, F.M. (2005), Analyzing Digital Terrain Attributes to Predict Soil Attributes for a Relatively Large Area, *Soil Science Society of America Journal*, vol. 69, pp. 1590-1599.

APPENDICES

Appendix A - Chapter 2

A.1 Neural Network PMML Code

The Neural Network models were produced in using the Data Mining Tool in Statistica (StatSoft, Inc., 2011). Below is a copy of the PMML code which the model used to produce D_b predictions for the 'Landscape variables only' ANN-A model (Table 2-4).


```
<?xml version="1.0" encoding="UTF-8"?>
<PMML version="3.0"><Header copyright="Copyright (c) StatSoft, Inc. All Rights
Reserved."><Application name="STATISTICA Automated Neural Networks (SANN)"
version="2.0"/></Header><DataDictionary numberOfFields="20"><DataField
name="Bulk_density" optype="continuous"/><DataField name="AAR"
optype="continuous"/><DataField name="AT0_ANNUAL"
optype="continuous"/><DataField name="FCD_MED" optype="continuous"/><DataField
name="PSMD" optype="continuous"/><DataField name="PT"
optype="continuous"/><DataField name="Aspect" optype="continuous"/><DataField
name="Curvature" optype="continuous"/><DataField name="Slope"
optype="continuous"/><DataField name="SWI" optype="continuous"/><DataField
name="STI" optype="continuous"/><DataField name="Elevation"
optype="continuous"/><DataField name="PM1" optype="categorical"><Value
value="Bb"/><Value value="Bg"/><Value value="Bh"/><Value value="Bo"/><Value
value="Bp"/><Value value="Cf"/><Value value="Da"/><Value value="Db"/><Value
value="Ea"/><Value value="Ef"/><Value value="Eg"/><Value value="Ei"/><Value
value="Fi"/><Value value="Fq"/><Value value="Fw"/><Value value="Fx"/><Value
value="Fy"/><Value value="Ga"/></DataField><DataField name="LANDUSE_SAMPLED"
optype="categorical"><Value value="AR"/><Value value="CO"/><Value
value="DC"/><Value value="FA"/><Value value="GC"/><Value value="HC"/><Value
value="LE"/><Value value="OR"/><Value value="OT"/><Value value="PG"/><Value
value="RC"/><Value value="RG"/><Value value="T?"/><Value
value="UG"/></DataField><DataField name="Pennock" optype="categorical"><Value
value="A"/><Value value="B"/><Value value="C"/><Value value="D"/><Value
value="E"/><Value value="F"/><Value value="G"/></DataField><DataField
name="Iwahashi" optype="categorical"><Value value="A"/><Value
value="B"/><Value value="C"/><Value value="D"/><Value value="E"/><Value
value="F"/><Value value="G"/><Value value="H"/></DataField><DataField
name="Great_group" optype="categorical"><Value value="Brown soils"/><Value
value="Ground-water gley soils"/><Value value="Man made soils"/><Value
value="Pelosols"/><Value value="Podzolic soils"/><Value value="Surface-water
gley soils"/></DataField><DataField name="Soil_association"
optype="categorical"><Value value="Cambic stagnogley soils"/><Value
value="Cambic stagnohumic gley soils"/><Value value="Ferritic brown
earths"/><Value value="Gleyic brown earths"/><Value value="Humo-ferric
podzols"/><Value value="Ironpan stagnopodzols"/><Value value="Man made
soils"/><Value value="Paleo-argillic stagnogley soils"/><Value value="Pelo-
alluvial gley soils"/><Value value="Pelo-stagnogley soils"/><Value
value="Stagnogleyic argillic brown earths"/><Value value="Typical argillic
brown earths"/><Value value="Typical argillic pelosols"/><Value value="Typical
brown alluvial soils"/><Value value="Typical brown calcareous earths"/><Value
value="Typical brown earths"/><Value value="Typical brown podzolic
soils"/><Value value="Typical brown sands"/><Value value="Typical calcareous
pelosols"/><Value value="Typical cambic gley soils"/><Value value="Typical
humic-sandy gley soils"/><Value value="Typical paleo-argillic brown
earths"/><Value value="Typical sandy gley soils"/><Value value="Typical
stagnogley soils"/></DataField><DataField name="RCS"
optype="categorical"><Value value="ARBR"/><Value value="ARSC"/><Value
```

value="ARSD"/><Value value="BREC"/><Value value="CONG"/><Value
value="DA"/><Value value="DOLO"/><Value value="FLIR"/><Value
value="FLMST"/><Value value="GNR"/><Value value="LMST"/><Value
value="LSMD"/><Value value="MDHA"/><Value value="MDLM"/><Value
value="MDSC"/><Value value="MDSD"/><Value value="MDSS"/><Value
value="MDST"/><Value value="PESST"/><Value value="SCON"/><Value
value="SCSM"/><Value value="SDLI"/><Value value="SDST"/><Value
value="SIMD"/><Value value="SISD"/><Value value="SLMDST"/><Value
value="SLST"/></DataField><DataField name="LEX" optype="categorical"><Value
value="AS"/><Value value="AW"/><Value value="BAN"/><Value value="BCMU"/><Value
value="BLCR"/><Value value="BLL"/><Value value="BMS"/><Value
value="BSG"/><Value value="CBRD"/><Value value="CDF"/><Value
value="CHAM"/><Value value="CHG"/><Value value="CLT"/><Value
value="CTM"/><Value value="DYS"/><Value value="ECL"/><Value
value="EDW"/><Value value="EN"/><Value value="ETM"/><Value value="GUN"/><Value
value="HA"/><Value value="HANS"/><Value value="HBR"/><Value
value="KDM"/><Value value="KHS"/><Value value="LES"/><Value
value="LLUS"/><Value value="LOS"/><Value value="MI"/><Value
value="MMG"/><Value value="MO"/><Value value="MOI"/><Value
value="MORRI"/><Value value="MRB"/><Value value="MVC"/><Value
value="NS"/><Value value="NTC"/><Value value="ONS"/><Value
value="OWSH"/><Value value="PET"/><Value value="PLCM"/><Value
value="PLD"/><Value value="PLWF"/><Value value="PMCM"/><Value
value="RG"/><Value value="RLS"/><Value value="RR"/><Value value="SASH"/><Value
value="SIM"/><Value value="SMG"/><Value value="SPPS"/><Value
value="TLM"/><Value value="TPSF"/><Value value="ULUS"/><Value
value="WBY"/><Value value="WCT"/><Value value="WDF"/><Value
value="WEL"/><Value value="WGF"/><Value value="WHM"/><Value
value="WIT"/><Value value="WOL"/><Value
value="WRS"/></DataField></DataDictionary><NeuralNetwork
modelName="A_Horizon_training_test_validation_inc_texture-_MLP 178-7-1"
functionName="regression"><MiningSchema><MiningField name="Bulk_density"
usageType="predicted"/><MiningField name="AAR" lowValue="596.000000"
highValue="1261.000000"/><MiningField name="AT0_ANNUAL" lowValue="2699.000000"
highValue="3830.000000"/><MiningField name="FCD_MED" lowValue="127.000000"
highValue="278.000000"/><MiningField name="PSMD" lowValue="61.000000"
highValue="251.000000"/><MiningField name="PT" lowValue="488.000000"
highValue="697.000000"/><MiningField name="Aspect" lowValue="-1.000000"
highValue="360.000000"/><MiningField name="Curvature" lowValue="-4.000000"
highValue="4.700000"/><MiningField name="Slope" lowValue="0.000000"
highValue="24.840000"/><MiningField name="SWI" lowValue="10.440000"
highValue="18.220000"/><MiningField name="STI" lowValue="-67.400000"
highValue="0.000000"/><MiningField name="Elevation" lowValue="12.700000"
highValue="403.500000"/><MiningField name="PM1"/><MiningField
name="LANDUSE_SAMPLED"/><MiningField name="Pennock"/><MiningField
name="Iwahashi"/><MiningField name="Great_group"/><MiningField
name="Soil_association"/><MiningField name="RCS"/><MiningField
name="LEX"/></MiningSchema><NeuralInputs numberOfInputs="178"><NeuralInput
id="0"><DerivedField><NormContinuous field="AAR" shift="-8.40875912408759e-
001" scale="1.45985401459854e-003"><LinearNorm orig="5.96000000000000e+002"
norm="0.000000"/><LinearNorm orig="1.26100000000000e+003"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="1"><DerivedField><NormContinuous field="AT0_ANNUAL" shift="-
2.38638373121132e+000" scale="8.84173297966401e-004"><LinearNorm
orig="2.69900000000000e+003" norm="0.000000"/><LinearNorm
orig="3.83000000000000e+003"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="2"><DerivedField><NormContinuous field="FCD_MED" shift="-
7.82051282051282e-001" scale="6.41025641025641e-003"><LinearNorm
orig="1.27000000000000e+002" norm="0.000000"/><LinearNorm
orig="2.78000000000000e+002"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="3"><DerivedField><NormContinuous field="PSMD" shift="-3.21052631578947e-
001" scale="5.26315789473684e-003"><LinearNorm orig="6.10000000000000e+001"
norm="0.000000"/><LinearNorm orig="2.51000000000000e+002"

```
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="4"><DerivedField><NormContinuous field="PT" shift="-2.33492822966507e+000"
scale="4.78468899521531e-003"><LinearNorm orig="4.88000000000000e+002"
norm="0.000000"/><LinearNorm orig="6.97000000000000e+002"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="5"><DerivedField><NormContinuous field="Aspect" shift="2.77008310249307e-
003" scale="2.77008310249307e-003"><LinearNorm orig="-1.00000000000000e+000"
norm="0.000000"/><LinearNorm orig="3.60000000000000e+002"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="6"><DerivedField><NormContinuous field="Curvature"
shift="4.59770114942529e-001" scale="1.14942528735632e-001"><LinearNorm
orig="-4.00000000000000e+000" norm="0.000000"/><LinearNorm
orig="4.70000000000000e+000"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="7"><DerivedField><NormContinuous field="Slope" shift="-
0.00000000000000e+000" scale="4.02576489533011e-002"><LinearNorm
orig="0.00000000000000e+000" norm="0.000000"/><LinearNorm
orig="2.48400000000000e+001"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="8"><DerivedField><NormContinuous field="SWI" shift="-
1.34190231362468e+000" scale="1.28534704370180e-001"><LinearNorm
orig="1.04400000000000e+001" norm="0.000000"/><LinearNorm
orig="1.82200000000000e+001"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="9"><DerivedField><NormContinuous field="STI" shift="1.00000000000000e+000"
scale="1.48367952522255e-002"><LinearNorm orig="-6.74000000000000e+001"
norm="0.000000"/><LinearNorm orig="0.00000000000000e+000"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="10"><DerivedField><NormContinuous field="Elevation" shift="-
2.28136882129278e-002" scale="2.53485424588086e-003"><LinearNorm
orig="1.27000000000000e+001" norm="0.000000"/><LinearNorm
orig="4.03500000000000e+002"
norm="1.000000"/></NormContinuous></DerivedField></NeuralInput><NeuralInput
id="11"><DerivedField><NormDiscrete field="PM1"
value="Bb"/></DerivedField></NeuralInput><NeuralInput
id="12"><DerivedField><NormDiscrete field="PM1"
value="Bg"/></DerivedField></NeuralInput><NeuralInput
id="13"><DerivedField><NormDiscrete field="PM1"
value="Bh"/></DerivedField></NeuralInput><NeuralInput
id="14"><DerivedField><NormDiscrete field="PM1"
value="Bo"/></DerivedField></NeuralInput><NeuralInput
id="15"><DerivedField><NormDiscrete field="PM1"
value="Bp"/></DerivedField></NeuralInput><NeuralInput
id="16"><DerivedField><NormDiscrete field="PM1"
value="Cf"/></DerivedField></NeuralInput><NeuralInput
id="17"><DerivedField><NormDiscrete field="PM1"
value="Da"/></DerivedField></NeuralInput><NeuralInput
id="18"><DerivedField><NormDiscrete field="PM1"
value="Db"/></DerivedField></NeuralInput><NeuralInput
id="19"><DerivedField><NormDiscrete field="PM1"
value="Ea"/></DerivedField></NeuralInput><NeuralInput
id="20"><DerivedField><NormDiscrete field="PM1"
value="Ef"/></DerivedField></NeuralInput><NeuralInput
id="21"><DerivedField><NormDiscrete field="PM1"
value="Eg"/></DerivedField></NeuralInput><NeuralInput
id="22"><DerivedField><NormDiscrete field="PM1"
value="Ei"/></DerivedField></NeuralInput><NeuralInput
id="23"><DerivedField><NormDiscrete field="PM1"
value="Fi"/></DerivedField></NeuralInput><NeuralInput
id="24"><DerivedField><NormDiscrete field="PM1"
value="Fq"/></DerivedField></NeuralInput><NeuralInput
id="25"><DerivedField><NormDiscrete field="PM1"
value="Fw"/></DerivedField></NeuralInput><NeuralInput
id="26"><DerivedField><NormDiscrete field="PM1"
value="Fx"/></DerivedField></NeuralInput><NeuralInput
```

```
id="27"><DerivedField><NormDiscrete field="PM1"
value="Fy"/></DerivedField></NeuralInput><NeuralInput
id="28"><DerivedField><NormDiscrete field="PM1"
value="Ga"/></DerivedField></NeuralInput><NeuralInput
id="29"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="AR"/></DerivedField></NeuralInput><NeuralInput
id="30"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="CO"/></DerivedField></NeuralInput><NeuralInput
id="31"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="DC"/></DerivedField></NeuralInput><NeuralInput
id="32"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="FA"/></DerivedField></NeuralInput><NeuralInput
id="33"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="GC"/></DerivedField></NeuralInput><NeuralInput
id="34"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="HC"/></DerivedField></NeuralInput><NeuralInput
id="35"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="LE"/></DerivedField></NeuralInput><NeuralInput
id="36"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="OR"/></DerivedField></NeuralInput><NeuralInput
id="37"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="OT"/></DerivedField></NeuralInput><NeuralInput
id="38"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="PG"/></DerivedField></NeuralInput><NeuralInput
id="39"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="RC"/></DerivedField></NeuralInput><NeuralInput
id="40"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="RG"/></DerivedField></NeuralInput><NeuralInput
id="41"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="T?"/></DerivedField></NeuralInput><NeuralInput
id="42"><DerivedField><NormDiscrete field="LANDUSE_SAMPLED"
value="UG"/></DerivedField></NeuralInput><NeuralInput
id="43"><DerivedField><NormDiscrete field="Pennock"
value="A"/></DerivedField></NeuralInput><NeuralInput
id="44"><DerivedField><NormDiscrete field="Pennock"
value="B"/></DerivedField></NeuralInput><NeuralInput
id="45"><DerivedField><NormDiscrete field="Pennock"
value="C"/></DerivedField></NeuralInput><NeuralInput
id="46"><DerivedField><NormDiscrete field="Pennock"
value="D"/></DerivedField></NeuralInput><NeuralInput
id="47"><DerivedField><NormDiscrete field="Pennock"
value="E"/></DerivedField></NeuralInput><NeuralInput
id="48"><DerivedField><NormDiscrete field="Pennock"
value="F"/></DerivedField></NeuralInput><NeuralInput
id="49"><DerivedField><NormDiscrete field="Pennock"
value="G"/></DerivedField></NeuralInput><NeuralInput
id="50"><DerivedField><NormDiscrete field="Iwahashi"
value="A"/></DerivedField></NeuralInput><NeuralInput
id="51"><DerivedField><NormDiscrete field="Iwahashi"
value="B"/></DerivedField></NeuralInput><NeuralInput
id="52"><DerivedField><NormDiscrete field="Iwahashi"
value="C"/></DerivedField></NeuralInput><NeuralInput
id="53"><DerivedField><NormDiscrete field="Iwahashi"
value="D"/></DerivedField></NeuralInput><NeuralInput
id="54"><DerivedField><NormDiscrete field="Iwahashi"
value="E"/></DerivedField></NeuralInput><NeuralInput
id="55"><DerivedField><NormDiscrete field="Iwahashi"
value="F"/></DerivedField></NeuralInput><NeuralInput
id="56"><DerivedField><NormDiscrete field="Iwahashi"
value="G"/></DerivedField></NeuralInput><NeuralInput
id="57"><DerivedField><NormDiscrete field="Iwahashi"
value="H"/></DerivedField></NeuralInput><NeuralInput
id="58"><DerivedField><NormDiscrete field="Great_group" value="Brown
soils"/></DerivedField></NeuralInput><NeuralInput
id="59"><DerivedField><NormDiscrete field="Great_group" value="Ground-water
```

```
gley soils"/></DerivedField></NeuralInput><NeuralInput
id="60"><DerivedField><NormDiscrete field="Great_group" value="Man made
soils"/></DerivedField></NeuralInput><NeuralInput
id="61"><DerivedField><NormDiscrete field="Great_group"
value="Pelosols"/></DerivedField></NeuralInput><NeuralInput
id="62"><DerivedField><NormDiscrete field="Great_group" value="Podzolic
soils"/></DerivedField></NeuralInput><NeuralInput
id="63"><DerivedField><NormDiscrete field="Great_group" value="Surface-water
gley soils"/></DerivedField></NeuralInput><NeuralInput
id="64"><DerivedField><NormDiscrete field="Soil_association" value="Cambic
stagnogley soils"/></DerivedField></NeuralInput><NeuralInput
id="65"><DerivedField><NormDiscrete field="Soil_association" value="Cambic
stagnohumic gley soils"/></DerivedField></NeuralInput><NeuralInput
id="66"><DerivedField><NormDiscrete field="Soil_association" value="Ferritic
brown earths"/></DerivedField></NeuralInput><NeuralInput
id="67"><DerivedField><NormDiscrete field="Soil_association" value="Gleyic
brown earths"/></DerivedField></NeuralInput><NeuralInput
id="68"><DerivedField><NormDiscrete field="Soil_association" value="Humo-
ferric podzols"/></DerivedField></NeuralInput><NeuralInput
id="69"><DerivedField><NormDiscrete field="Soil_association" value="Ironpan
stagnopodzols"/></DerivedField></NeuralInput><NeuralInput
id="70"><DerivedField><NormDiscrete field="Soil_association" value="Man made
soils"/></DerivedField></NeuralInput><NeuralInput
id="71"><DerivedField><NormDiscrete field="Soil_association" value="Paleo-
argillic stagnogley soils"/></DerivedField></NeuralInput><NeuralInput
id="72"><DerivedField><NormDiscrete field="Soil_association" value="Pelo-
alluvial gley soils"/></DerivedField></NeuralInput><NeuralInput
id="73"><DerivedField><NormDiscrete field="Soil_association" value="Pelo-
stagnogley soils"/></DerivedField></NeuralInput><NeuralInput
id="74"><DerivedField><NormDiscrete field="Soil_association"
value="Stagnogleyic argillic brown
earths"/></DerivedField></NeuralInput><NeuralInput
id="75"><DerivedField><NormDiscrete field="Soil_association" value="Typical
argillic brown earths"/></DerivedField></NeuralInput><NeuralInput
id="76"><DerivedField><NormDiscrete field="Soil_association" value="Typical
argillic pelosols"/></DerivedField></NeuralInput><NeuralInput
id="77"><DerivedField><NormDiscrete field="Soil_association" value="Typical
brown alluvial soils"/></DerivedField></NeuralInput><NeuralInput
id="78"><DerivedField><NormDiscrete field="Soil_association" value="Typical
brown calcareous earths"/></DerivedField></NeuralInput><NeuralInput
id="79"><DerivedField><NormDiscrete field="Soil_association" value="Typical
brown earths"/></DerivedField></NeuralInput><NeuralInput
id="80"><DerivedField><NormDiscrete field="Soil_association" value="Typical
brown podzolic soils"/></DerivedField></NeuralInput><NeuralInput
id="81"><DerivedField><NormDiscrete field="Soil_association" value="Typical
brown sands"/></DerivedField></NeuralInput><NeuralInput
id="82"><DerivedField><NormDiscrete field="Soil_association" value="Typical
calcareous pelosols"/></DerivedField></NeuralInput><NeuralInput
id="83"><DerivedField><NormDiscrete field="Soil_association" value="Typical
cambic gley soils"/></DerivedField></NeuralInput><NeuralInput
id="84"><DerivedField><NormDiscrete field="Soil_association" value="Typical
humic-sandy gley soils"/></DerivedField></NeuralInput><NeuralInput
id="85"><DerivedField><NormDiscrete field="Soil_association" value="Typical
paleo-argillic brown earths"/></DerivedField></NeuralInput><NeuralInput
id="86"><DerivedField><NormDiscrete field="Soil_association" value="Typical
sandy gley soils"/></DerivedField></NeuralInput><NeuralInput
id="87"><DerivedField><NormDiscrete field="Soil_association" value="Typical
stagnogley soils"/></DerivedField></NeuralInput><NeuralInput
id="88"><DerivedField><NormDiscrete field="RCS"
value="ARBR"/></DerivedField></NeuralInput><NeuralInput
id="89"><DerivedField><NormDiscrete field="RCS"
value="ARSC"/></DerivedField></NeuralInput><NeuralInput
id="90"><DerivedField><NormDiscrete field="RCS"
value="ARSD"/></DerivedField></NeuralInput><NeuralInput
id="91"><DerivedField><NormDiscrete field="RCS"
```

```
value="BREC"/></DerivedField></NeuralInput><NeuralInput
id="92"><DerivedField><NormDiscrete field="RCS"
value="CONG"/></DerivedField></NeuralInput><NeuralInput
id="93"><DerivedField><NormDiscrete field="RCS"
value="DA"/></DerivedField></NeuralInput><NeuralInput
id="94"><DerivedField><NormDiscrete field="RCS"
value="DOLO"/></DerivedField></NeuralInput><NeuralInput
id="95"><DerivedField><NormDiscrete field="RCS"
value="FLIR"/></DerivedField></NeuralInput><NeuralInput
id="96"><DerivedField><NormDiscrete field="RCS"
value="FLMST"/></DerivedField></NeuralInput><NeuralInput
id="97"><DerivedField><NormDiscrete field="RCS"
value="GNR"/></DerivedField></NeuralInput><NeuralInput
id="98"><DerivedField><NormDiscrete field="RCS"
value="LMST"/></DerivedField></NeuralInput><NeuralInput
id="99"><DerivedField><NormDiscrete field="RCS"
value="LSMD"/></DerivedField></NeuralInput><NeuralInput
id="100"><DerivedField><NormDiscrete field="RCS"
value="MDHA"/></DerivedField></NeuralInput><NeuralInput
id="101"><DerivedField><NormDiscrete field="RCS"
value="MDLM"/></DerivedField></NeuralInput><NeuralInput
id="102"><DerivedField><NormDiscrete field="RCS"
value="MDSC"/></DerivedField></NeuralInput><NeuralInput
id="103"><DerivedField><NormDiscrete field="RCS"
value="MDSD"/></DerivedField></NeuralInput><NeuralInput
id="104"><DerivedField><NormDiscrete field="RCS"
value="MDSS"/></DerivedField></NeuralInput><NeuralInput
id="105"><DerivedField><NormDiscrete field="RCS"
value="MDST"/></DerivedField></NeuralInput><NeuralInput
id="106"><DerivedField><NormDiscrete field="RCS"
value="PESST"/></DerivedField></NeuralInput><NeuralInput
id="107"><DerivedField><NormDiscrete field="RCS"
value="SCON"/></DerivedField></NeuralInput><NeuralInput
id="108"><DerivedField><NormDiscrete field="RCS"
value="SCSM"/></DerivedField></NeuralInput><NeuralInput
id="109"><DerivedField><NormDiscrete field="RCS"
value="SDLI"/></DerivedField></NeuralInput><NeuralInput
id="110"><DerivedField><NormDiscrete field="RCS"
value="SDST"/></DerivedField></NeuralInput><NeuralInput
id="111"><DerivedField><NormDiscrete field="RCS"
value="SIMD"/></DerivedField></NeuralInput><NeuralInput
id="112"><DerivedField><NormDiscrete field="RCS"
value="SISD"/></DerivedField></NeuralInput><NeuralInput
id="113"><DerivedField><NormDiscrete field="RCS"
value="SLMDST"/></DerivedField></NeuralInput><NeuralInput
id="114"><DerivedField><NormDiscrete field="RCS"
value="SLST"/></DerivedField></NeuralInput><NeuralInput
id="115"><DerivedField><NormDiscrete field="LEX"
value="AS"/></DerivedField></NeuralInput><NeuralInput
id="116"><DerivedField><NormDiscrete field="LEX"
value="AW"/></DerivedField></NeuralInput><NeuralInput
id="117"><DerivedField><NormDiscrete field="LEX"
value="BAN"/></DerivedField></NeuralInput><NeuralInput
id="118"><DerivedField><NormDiscrete field="LEX"
value="BCMU"/></DerivedField></NeuralInput><NeuralInput
id="119"><DerivedField><NormDiscrete field="LEX"
value="BLCR"/></DerivedField></NeuralInput><NeuralInput
id="120"><DerivedField><NormDiscrete field="LEX"
value="BLL"/></DerivedField></NeuralInput><NeuralInput
id="121"><DerivedField><NormDiscrete field="LEX"
value="BMS"/></DerivedField></NeuralInput><NeuralInput
id="122"><DerivedField><NormDiscrete field="LEX"
value="BSG"/></DerivedField></NeuralInput><NeuralInput
id="123"><DerivedField><NormDiscrete field="LEX"
value="CBRD"/></DerivedField></NeuralInput><NeuralInput
```

id="124"><DerivedField><NormDiscrete field="LEX"
value="CDF"/></DerivedField></NeuralInput><NeuralInput
id="125"><DerivedField><NormDiscrete field="LEX"
value="CHAM"/></DerivedField></NeuralInput><NeuralInput
id="126"><DerivedField><NormDiscrete field="LEX"
value="CHG"/></DerivedField></NeuralInput><NeuralInput
id="127"><DerivedField><NormDiscrete field="LEX"
value="CLT"/></DerivedField></NeuralInput><NeuralInput
id="128"><DerivedField><NormDiscrete field="LEX"
value="CTM"/></DerivedField></NeuralInput><NeuralInput
id="129"><DerivedField><NormDiscrete field="LEX"
value="DYS"/></DerivedField></NeuralInput><NeuralInput
id="130"><DerivedField><NormDiscrete field="LEX"
value="ECL"/></DerivedField></NeuralInput><NeuralInput
id="131"><DerivedField><NormDiscrete field="LEX"
value="EDW"/></DerivedField></NeuralInput><NeuralInput
id="132"><DerivedField><NormDiscrete field="LEX"
value="EN"/></DerivedField></NeuralInput><NeuralInput
id="133"><DerivedField><NormDiscrete field="LEX"
value="ETM"/></DerivedField></NeuralInput><NeuralInput
id="134"><DerivedField><NormDiscrete field="LEX"
value="GUN"/></DerivedField></NeuralInput><NeuralInput
id="135"><DerivedField><NormDiscrete field="LEX"
value="HA"/></DerivedField></NeuralInput><NeuralInput
id="136"><DerivedField><NormDiscrete field="LEX"
value="HANS"/></DerivedField></NeuralInput><NeuralInput
id="137"><DerivedField><NormDiscrete field="LEX"
value="HBR"/></DerivedField></NeuralInput><NeuralInput
id="138"><DerivedField><NormDiscrete field="LEX"
value="KDM"/></DerivedField></NeuralInput><NeuralInput
id="139"><DerivedField><NormDiscrete field="LEX"
value="KHS"/></DerivedField></NeuralInput><NeuralInput
id="140"><DerivedField><NormDiscrete field="LEX"
value="LES"/></DerivedField></NeuralInput><NeuralInput
id="141"><DerivedField><NormDiscrete field="LEX"
value="LLUS"/></DerivedField></NeuralInput><NeuralInput
id="142"><DerivedField><NormDiscrete field="LEX"
value="LOS"/></DerivedField></NeuralInput><NeuralInput
id="143"><DerivedField><NormDiscrete field="LEX"
value="MI"/></DerivedField></NeuralInput><NeuralInput
id="144"><DerivedField><NormDiscrete field="LEX"
value="MMG"/></DerivedField></NeuralInput><NeuralInput
id="145"><DerivedField><NormDiscrete field="LEX"
value="MO"/></DerivedField></NeuralInput><NeuralInput
id="146"><DerivedField><NormDiscrete field="LEX"
value="MOI"/></DerivedField></NeuralInput><NeuralInput
id="147"><DerivedField><NormDiscrete field="LEX"
value="MORRI"/></DerivedField></NeuralInput><NeuralInput
id="148"><DerivedField><NormDiscrete field="LEX"
value="MRB"/></DerivedField></NeuralInput><NeuralInput
id="149"><DerivedField><NormDiscrete field="LEX"
value="MVC"/></DerivedField></NeuralInput><NeuralInput
id="150"><DerivedField><NormDiscrete field="LEX"
value="NS"/></DerivedField></NeuralInput><NeuralInput
id="151"><DerivedField><NormDiscrete field="LEX"
value="NTC"/></DerivedField></NeuralInput><NeuralInput
id="152"><DerivedField><NormDiscrete field="LEX"
value="ONS"/></DerivedField></NeuralInput><NeuralInput
id="153"><DerivedField><NormDiscrete field="LEX"
value="OWSH"/></DerivedField></NeuralInput><NeuralInput
id="154"><DerivedField><NormDiscrete field="LEX"
value="PET"/></DerivedField></NeuralInput><NeuralInput
id="155"><DerivedField><NormDiscrete field="LEX"
value="PLCM"/></DerivedField></NeuralInput><NeuralInput
id="156"><DerivedField><NormDiscrete field="LEX"

value="PLD"/></DerivedField></NeuralInput><NeuralInput
id="157"><DerivedField><NormDiscrete field="LEX"
value="PLWF"/></DerivedField></NeuralInput><NeuralInput
id="158"><DerivedField><NormDiscrete field="LEX"
value="PMCM"/></DerivedField></NeuralInput><NeuralInput
id="159"><DerivedField><NormDiscrete field="LEX"
value="RG"/></DerivedField></NeuralInput><NeuralInput
id="160"><DerivedField><NormDiscrete field="LEX"
value="RLS"/></DerivedField></NeuralInput><NeuralInput
id="161"><DerivedField><NormDiscrete field="LEX"
value="RR"/></DerivedField></NeuralInput><NeuralInput
id="162"><DerivedField><NormDiscrete field="LEX"
value="SASH"/></DerivedField></NeuralInput><NeuralInput
id="163"><DerivedField><NormDiscrete field="LEX"
value="SIM"/></DerivedField></NeuralInput><NeuralInput
id="164"><DerivedField><NormDiscrete field="LEX"
value="SMG"/></DerivedField></NeuralInput><NeuralInput
id="165"><DerivedField><NormDiscrete field="LEX"
value="SPPS"/></DerivedField></NeuralInput><NeuralInput
id="166"><DerivedField><NormDiscrete field="LEX"
value="TLM"/></DerivedField></NeuralInput><NeuralInput
id="167"><DerivedField><NormDiscrete field="LEX"
value="TPSF"/></DerivedField></NeuralInput><NeuralInput
id="168"><DerivedField><NormDiscrete field="LEX"
value="ULUS"/></DerivedField></NeuralInput><NeuralInput
id="169"><DerivedField><NormDiscrete field="LEX"
value="WBY"/></DerivedField></NeuralInput><NeuralInput
id="170"><DerivedField><NormDiscrete field="LEX"
value="WCI"/></DerivedField></NeuralInput><NeuralInput
id="171"><DerivedField><NormDiscrete field="LEX"
value="WDF"/></DerivedField></NeuralInput><NeuralInput
id="172"><DerivedField><NormDiscrete field="LEX"
value="WEL"/></DerivedField></NeuralInput><NeuralInput
id="173"><DerivedField><NormDiscrete field="LEX"
value="WGF"/></DerivedField></NeuralInput><NeuralInput
id="174"><DerivedField><NormDiscrete field="LEX"
value="WHM"/></DerivedField></NeuralInput><NeuralInput
id="175"><DerivedField><NormDiscrete field="LEX"
value="WIT"/></DerivedField></NeuralInput><NeuralInput
id="176"><DerivedField><NormDiscrete field="LEX"
value="WOL"/></DerivedField></NeuralInput><NeuralInput
id="177"><DerivedField><NormDiscrete field="LEX"
value="WRS"/></DerivedField></NeuralInput></NeuralInputs><NeuralLayer
numberOfNeurons="7" activationFunction="sine"><Neuron id="178" bias="-
2.40991699473977e-001"><Con from="0" weight="-3.19744398171522e-001"/><Con
from="1" weight="3.75542699514308e-002"/><Con from="2" weight="-
3.60326625177779e-001"/><Con from="3" weight="9.28211209211258e-002"/><Con
from="4" weight="1.71004460823147e-002"/><Con from="5" weight="-
1.10830363218681e-001"/><Con from="6" weight="-5.18537812410757e-002"/><Con
from="7" weight="-1.00535417998738e-001"/><Con from="8" weight="-
5.14993706428754e-002"/><Con from="9" weight="-2.09522013494974e-001"/><Con
from="10" weight="-3.16250548928865e-001"/><Con from="11" weight="-
2.93888328556738e-002"/><Con from="12" weight="-8.74828632207029e-003"/><Con
from="13" weight="-4.40290325581359e-002"/><Con from="14"
weight="2.75739330091938e-002"/><Con from="15" weight="3.60961079886057e-
003"/><Con from="16" weight="8.73358271496331e-003"/><Con from="17" weight="-
8.10451978403517e-004"/><Con from="18" weight="3.07005031532661e-003"/><Con
from="19" weight="-2.22733588093558e-001"/><Con from="20" weight="-
2.53756779695223e-002"/><Con from="21" weight="7.08919753098209e-002"/><Con
from="22" weight="8.87730537321409e-002"/><Con from="23" weight="-
2.94186063700435e-001"/><Con from="24" weight="8.68447179407204e-002"/><Con
from="25" weight="-2.04559572906588e-003"/><Con from="26"
weight="2.65380596488908e-002"/><Con from="27" weight="4.17263648394817e-
003"/><Con from="28" weight="4.99443249138638e-002"/><Con from="29"
weight="4.64281371255944e-001"/><Con from="30" weight="-3.85786111238977e-

002"/><Con from="31" weight="-1.06788515019081e-001"/><Con from="32"
weight="6.92400108269857e-002"/><Con from="33" weight="2.70562422367116e-
002"/><Con from="34" weight="8.93739249264446e-005"/><Con from="35"
weight="1.73421413856698e-001"/><Con from="36" weight="9.89068215340010e-
003"/><Con from="37" weight="-1.27143430094357e-001"/><Con from="38" weight="-
6.01527809391728e-001"/><Con from="39" weight="4.26476221036770e-003"/><Con
from="40" weight="-5.22622806064451e-002"/><Con from="41" weight="-
2.93932604945807e-002"/><Con from="42" weight="-4.29964772223374e-002"/><Con
from="43" weight="-1.90087461845023e-001"/><Con from="44" weight="-
8.49514126951984e-002"/><Con from="45" weight="-1.80811285316242e-002"/><Con
from="46" weight="2.38558182421003e-003"/><Con from="47" weight="-
6.71669997137265e-002"/><Con from="48" weight="-4.48103940396652e-002"/><Con
from="49" weight="1.21288512974495e-001"/><Con from="50" weight="-
1.62534155363281e-001"/><Con from="51" weight="-1.35905927938732e-001"/><Con
from="52" weight="3.40129659142981e-002"/><Con from="53"
weight="7.42983709134241e-003"/><Con from="54" weight="1.47350062650179e-
001"/><Con from="55" weight="-6.24847514249128e-002"/><Con from="56" weight="-
7.10964344355658e-002"/><Con from="57" weight="1.17146716356055e-002"/><Con
from="58" weight="3.52695767560160e-001"/><Con from="59" weight="-
2.16522403520146e-001"/><Con from="60" weight="2.94829463568496e-002"/><Con
from="61" weight="-3.75204308746357e-002"/><Con from="62" weight="-
7.56185346918680e-002"/><Con from="63" weight="-3.12784396630649e-001"/><Con
from="64" weight="-3.58543810007835e-002"/><Con from="65" weight="-
1.31823220344819e-001"/><Con from="66" weight="1.16693251210724e-002"/><Con
from="67" weight="-1.06333904500088e-002"/><Con from="68" weight="-
2.19291169870247e-002"/><Con from="69" weight="-4.26875597462096e-002"/><Con
from="70" weight="3.05241215791605e-002"/><Con from="71"
weight="1.60729044905058e-002"/><Con from="72" weight="-2.32597647019807e-
001"/><Con from="73" weight="-1.31692906000804e-001"/><Con from="74"
weight="4.87003162588653e-002"/><Con from="75" weight="2.46858224748914e-
002"/><Con from="76" weight="-3.61491424379937e-002"/><Con from="77" weight="-
1.20854508848105e-002"/><Con from="78" weight="-1.62948116707071e-002"/><Con
from="79" weight="1.86054552375822e-001"/><Con from="80" weight="-
2.94944312138975e-002"/><Con from="81" weight="1.19071238680345e-001"/><Con
from="82" weight="1.17434912942192e-002"/><Con from="83"
weight="1.28314816403548e-002"/><Con from="84" weight="-2.22090053237166e-
002"/><Con from="85" weight="1.78567367408543e-003"/><Con from="86"
weight="9.87419736583796e-003"/><Con from="87" weight="-2.84798832605155e-
002"/><Con from="88" weight="-3.06721767899458e-003"/><Con from="89"
weight="1.48332406224107e-002"/><Con from="90" weight="1.93697847163243e-
002"/><Con from="91" weight="-3.42189855600069e-002"/><Con from="92" weight="-
7.27505675635385e-004"/><Con from="93" weight="-3.35176518692317e-002"/><Con
from="94" weight="-1.02671883861055e-002"/><Con from="95" weight="-
8.50606759676944e-003"/><Con from="96" weight="-1.19435299541823e-002"/><Con
from="97" weight="-1.07239959225657e-002"/><Con from="98" weight="-
1.91234476416991e-002"/><Con from="99" weight="-1.06742771100627e-001"/><Con
from="100" weight="-9.35055034985014e-002"/><Con from="101"
weight="3.30578021291573e-003"/><Con from="102" weight="-9.11192170128781e-
004"/><Con from="103" weight="-5.19142494860016e-004"/><Con from="104"
weight="-2.22542377625322e-001"/><Con from="105" weight="5.92373063443886e-
002"/><Con from="106" weight="8.87577006646423e-002"/><Con from="107"
weight="-1.02834424594143e-002"/><Con from="108" weight="2.30885907564652e-
002"/><Con from="109" weight="2.21386592615363e-002"/><Con from="110"
weight="8.57586952002656e-002"/><Con from="111" weight="-7.03895669446948e-
003"/><Con from="112" weight="4.12382227919547e-002"/><Con from="113"
weight="-1.04595002057273e-002"/><Con from="114" weight="-5.78170171505148e-
004"/><Con from="115" weight="2.07538562652741e-002"/><Con from="116"
weight="-1.03448159891479e-002"/><Con from="117" weight="-8.56314311552819e-
002"/><Con from="118" weight="-4.47165595597135e-003"/><Con from="119"
weight="-2.32330165143737e-002"/><Con from="120" weight="-1.71901712276716e-
003"/><Con from="121" weight="1.33074429287474e-001"/><Con from="122"
weight="-8.08978158874028e-002"/><Con from="123" weight="2.40995310476442e-
004"/><Con from="124" weight="-1.54768746078594e-003"/><Con from="125"
weight="4.60076794922385e-002"/><Con from="126" weight="-1.02691575099565e-
002"/><Con from="127" weight="9.99923876681126e-003"/><Con from="128"

weight="-4.06317371411721e-003"/><Con from="129" weight="-3.48807278921896e-002"/><Con from="130" weight="7.21149558386236e-003"/><Con from="131" weight="1.55844156783866e-002"/><Con from="132" weight="2.32238687234963e-002"/><Con from="133" weight="-3.50173063111306e-003"/><Con from="134" weight="2.58213418426843e-002"/><Con from="135" weight="2.48870443899163e-002"/><Con from="136" weight="1.56528011727641e-002"/><Con from="137" weight="-4.45715565434198e-002"/><Con from="138" weight="-5.58885912280768e-002"/><Con from="139" weight="4.51283803181740e-002"/><Con from="140" weight="1.19498289112184e-003"/><Con from="141" weight="7.97332075959961e-003"/><Con from="142" weight="-3.96254047845031e-002"/><Con from="143" weight="4.21301117236113e-003"/><Con from="144" weight="1.83975436074257e-002"/><Con from="145" weight="-1.65365554582585e-002"/><Con from="146" weight="3.43786653682270e-003"/><Con from="147" weight="-1.99260675934300e-001"/><Con from="148" weight="-6.63045896703467e-004"/><Con from="149" weight="-3.25665187044164e-002"/><Con from="150" weight="2.25864879374544e-002"/><Con from="151" weight="4.67839545740620e-002"/><Con from="152" weight="-1.35656238503549e-002"/><Con from="153" weight="-2.77840717022024e-002"/><Con from="154" weight="1.00589536310417e-002"/><Con from="155" weight="-2.61933903476780e-003"/><Con from="156" weight="-3.86544821405607e-002"/><Con from="157" weight="1.55826669329876e-003"/><Con from="158" weight="2.49643486842652e-002"/><Con from="159" weight="-2.14394008952974e-003"/><Con from="160" weight="2.37087197541336e-002"/><Con from="161" weight="7.96184715809298e-003"/><Con from="162" weight="1.89997518431238e-002"/><Con from="163" weight="-5.62754170807755e-002"/><Con from="164" weight="1.48469510951188e-002"/><Con from="165" weight="6.30193311545171e-003"/><Con from="166" weight="2.33364635745251e-004"/><Con from="167" weight="2.60367290669257e-002"/><Con from="168" weight="5.41591837603538e-003"/><Con from="169" weight="-5.37669008065817e-002"/><Con from="170" weight="-2.88611259800487e-002"/><Con from="171" weight="-1.11894073323355e-001"/><Con from="172" weight="-9.44814288734379e-004"/><Con from="173" weight="1.94486479460276e-002"/><Con from="174" weight="-5.57317146138607e-003"/><Con from="175" weight="1.62833368316886e-002"/><Con from="176" weight="3.38097282290032e-003"/><Con from="177" weight="3.01758959041173e-002"/></Neuron><Neuron id="179" bias="-1.36499494933831e-001"><Con from="0" weight="-1.03006468243096e-001"/><Con from="1" weight="-6.13217774703352e-003"/><Con from="2" weight="-1.06271426714311e-001"/><Con from="3" weight="-9.83719830720455e-003"/><Con from="4" weight="-2.93595740486591e-002"/><Con from="5" weight="-4.34536027485683e-002"/><Con from="6" weight="-3.53253195213743e-002"/><Con from="7" weight="-3.43664731956815e-002"/><Con from="8" weight="-3.06091132177443e-002"/><Con from="9" weight="-1.17231218549438e-001"/><Con from="10" weight="-1.12986101954477e-001"/><Con from="11" weight="-1.84138087789964e-002"/><Con from="12" weight="-1.04219349995456e-002"/><Con from="13" weight="-2.28232577577206e-003"/><Con from="14" weight="1.36880938062752e-002"/><Con from="15" weight="-4.57010766096920e-003"/><Con from="16" weight="-5.24107214620644e-003"/><Con from="17" weight="-9.40630450223553e-004"/><Con from="18" weight="-1.48261576447252e-002"/><Con from="19" weight="-5.78036760440995e-002"/><Con from="20" weight="-1.39425286138003e-002"/><Con from="21" weight="1.69992238225749e-002"/><Con from="22" weight="-4.17866115554011e-003"/><Con from="23" weight="-1.34979618293781e-001"/><Con from="24" weight="4.68500517091108e-002"/><Con from="25" weight="5.07988411250385e-003"/><Con from="26" weight="3.37697283540443e-003"/><Con from="27" weight="-2.86245052612371e-003"/><Con from="28" weight="1.59343477470823e-002"/><Con from="29" weight="9.36671794505733e-002"/><Con from="30" weight="-4.81092598280853e-003"/><Con from="31" weight="-6.29958607847579e-002"/><Con from="32" weight="2.28163331664246e-002"/><Con from="33" weight="1.55824591933671e-002"/><Con from="34" weight="-1.04332692625742e-003"/><Con from="35" weight="5.81364111904801e-002"/><Con from="36" weight="2.98000776347569e-003"/><Con from="37" weight="-4.49303688774769e-002"/><Con from="38" weight="-1.40554747218318e-001"/><Con from="39" weight="-5.48658489894259e-003"/><Con from="40" weight="-2.71618746980129e-002"/><Con from="41" weight="-2.17621437034146e-002"/><Con from="42" weight="-1.57365193395975e-002"/><Con from="43" weight="-3.64053590391622e-002"/><Con from="44" weight="-3.31593185234254e-002"/><Con from="45" weight="-2.82042915756902e-003"/><Con from="46" weight="-9.85113925626918e-003"/><Con

from="47" weight="-2.93236246951254e-002"/><Con from="48" weight="-
1.31822087823713e-002"/><Con from="49" weight="3.42735274528069e-002"/><Con
from="50" weight="-7.79533656814248e-002"/><Con from="51" weight="-
6.08127703474468e-002"/><Con from="52" weight="4.93017001422834e-003"/><Con
from="53" weight="9.86161119691363e-003"/><Con from="54"
weight="1.99178238151726e-002"/><Con from="55" weight="-9.81722283330762e-
003"/><Con from="56" weight="-2.25732403663899e-002"/><Con from="57" weight="-
8.65946880124998e-003"/><Con from="58" weight="1.00451642391485e-001"/><Con
from="59" weight="-4.45948364840843e-002"/><Con from="60"
weight="2.13825797538838e-002"/><Con from="61" weight="-1.62605599708451e-
002"/><Con from="62" weight="-3.35647484386138e-002"/><Con from="63" weight="-
1.35264987869974e-001"/><Con from="64" weight="-6.09101344886576e-003"/><Con
from="65" weight="-4.07729015581554e-002"/><Con from="66" weight="-
6.36339632088771e-003"/><Con from="67" weight="8.60944268710314e-003"/><Con
from="68" weight="-2.25081821424945e-003"/><Con from="69" weight="-
8.23268043954572e-003"/><Con from="70" weight="1.17067536963372e-002"/><Con
from="71" weight="9.59024788670108e-003"/><Con from="72" weight="-
6.28600600248283e-002"/><Con from="73" weight="-4.65125638559948e-002"/><Con
from="74" weight="3.90021662124076e-004"/><Con from="75"
weight="1.47032490496911e-003"/><Con from="76" weight="-1.98705948945962e-
002"/><Con from="77" weight="6.86041796180855e-003"/><Con from="78" weight="-
1.15114172252676e-003"/><Con from="79" weight="5.46486164415067e-002"/><Con
from="80" weight="-6.03300801669623e-003"/><Con from="81"
weight="3.77228715888128e-002"/><Con from="82" weight="-1.93381620734236e-
003"/><Con from="83" weight="-6.58757308655219e-004"/><Con from="84"
weight="1.50952268874273e-003"/><Con from="85" weight="3.44782305589251e-
004"/><Con from="86" weight="2.18622038376202e-002"/><Con from="87" weight="-
6.97906439835337e-002"/><Con from="88" weight="9.53037924599166e-003"/><Con
from="89" weight="1.19319618025185e-002"/><Con from="90"
weight="1.94509379112436e-002"/><Con from="91" weight="-8.00194411204311e-
003"/><Con from="92" weight="-6.34735985068730e-003"/><Con from="93" weight="-
5.63010188522123e-003"/><Con from="94" weight="-5.24126689530909e-003"/><Con
from="95" weight="-7.33701626093345e-003"/><Con from="96"
weight="9.99862377456993e-003"/><Con from="97" weight="-5.65831817397114e-
003"/><Con from="98" weight="1.73551752859782e-004"/><Con from="99" weight="-
5.75152133909560e-002"/><Con from="100" weight="-4.20314966298410e-002"/><Con
from="101" weight="-3.86810034857465e-003"/><Con from="102" weight="-
1.32835794011207e-003"/><Con from="103" weight="1.20537785731317e-002"/><Con
from="104" weight="-7.15126726779197e-002"/><Con from="105" weight="-
1.16905508122088e-002"/><Con from="106" weight="8.90735698577088e-003"/><Con
from="107" weight="-8.22941810048034e-003"/><Con from="108" weight="-
4.38806493900845e-003"/><Con from="109" weight="5.80242284785488e-003"/><Con
from="110" weight="1.29968976863369e-002"/><Con from="111" weight="-
9.62349002370210e-003"/><Con from="112" weight="2.28109840700779e-002"/><Con
from="113" weight="1.88828669130019e-003"/><Con from="114" weight="-
9.40224927851112e-003"/><Con from="115" weight="3.59950500394270e-003"/><Con
from="116" weight="2.48034419311079e-003"/><Con from="117" weight="-
3.24046670107779e-002"/><Con from="118" weight="-8.31518043240957e-003"/><Con
from="119" weight="-1.25214341732396e-002"/><Con from="120" weight="-
6.64975042479663e-003"/><Con from="121" weight="3.81744109370127e-002"/><Con
from="122" weight="-2.16247523361145e-002"/><Con from="123" weight="-
5.16307338195893e-003"/><Con from="124" weight="1.07579955276756e-002"/><Con
from="125" weight="2.02272749098795e-002"/><Con from="126"
weight="1.66792135502905e-003"/><Con from="127" weight="1.75398983494402e-
003"/><Con from="128" weight="3.70624239327226e-004"/><Con from="129"
weight="-1.82419057671438e-002"/><Con from="130" weight="1.28552007127563e-
003"/><Con from="131" weight="4.63280401457214e-004"/><Con from="132"
weight="6.51873443683740e-003"/><Con from="133" weight="1.15332109139486e-
002"/><Con from="134" weight="7.85478095742881e-003"/><Con from="135"
weight="6.61337608442104e-003"/><Con from="136" weight="8.95387626183638e-
003"/><Con from="137" weight="-1.22164925931172e-002"/><Con from="138"
weight="-2.09820530447677e-002"/><Con from="139" weight="1.30868030375949e-
002"/><Con from="140" weight="8.93593503804513e-003"/><Con from="141"
weight="7.20265139043550e-003"/><Con from="142" weight="-1.23977850447761e-
002"/><Con from="143" weight="6.73210597888335e-003"/><Con from="144"

weight="1.71611366154152e-003"/><Con from="145" weight="-5.78540778084241e-003"/><Con from="146" weight="5.84861333327544e-003"/><Con from="147" weight="-6.16302811908397e-002"/><Con from="148" weight="-6.12088323850527e-003"/><Con from="149" weight="-1.61877045814985e-002"/><Con from="150" weight="8.71145220864360e-003"/><Con from="151" weight="8.50166552576982e-003"/><Con from="152" weight="-6.08787449341462e-003"/><Con from="153" weight="-3.31673839882240e-002"/><Con from="154" weight="-5.13986281523586e-004"/><Con from="155" weight="-3.33209001082574e-003"/><Con from="156" weight="-1.63725026989978e-002"/><Con from="157" weight="8.79166445971384e-003"/><Con from="158" weight="6.09331631638568e-003"/><Con from="159" weight="4.83631088168619e-003"/><Con from="160" weight="5.35115232647369e-003"/><Con from="161" weight="4.72395813218940e-004"/><Con from="162" weight="5.81454424550465e-003"/><Con from="163" weight="-5.97213049899585e-003"/><Con from="164" weight="-6.48722121694041e-003"/><Con from="165" weight="-7.84473735610470e-003"/><Con from="166" weight="-4.60746010741651e-003"/><Con from="167" weight="5.03148541542238e-003"/><Con from="168" weight="-3.66821241668084e-004"/><Con from="169" weight="-2.61735497499008e-002"/><Con from="170" weight="-5.48480955728684e-003"/><Con from="171" weight="-2.14388064689828e-002"/><Con from="172" weight="-8.62057155784239e-003"/><Con from="173" weight="1.19475950673327e-002"/><Con from="174" weight="-2.95930150300421e-003"/><Con from="175" weight="1.29156150527096e-002"/><Con from="176" weight="-3.49041614744724e-003"/><Con from="177" weight="1.31515058388664e-002"/></Neuron><Neuron id="180" bias="-5.64844487512015e-003"><Con from="0" weight="-1.67199807408107e-001"/><Con from="1" weight="2.89578762919741e-002"/><Con from="2" weight="-1.69825556008132e-001"/><Con from="3" weight="5.22992764317437e-002"/><Con from="4" weight="3.49982160102747e-002"/><Con from="5" weight="5.55078958592696e-002"/><Con from="6" weight="4.29848671949198e-002"/><Con from="7" weight="-2.46945297345568e-002"/><Con from="8" weight="8.83049625562280e-002"/><Con from="9" weight="-1.25608205669523e-002"/><Con from="10" weight="-1.21412769617922e-001"/><Con from="11" weight="-2.22229470639398e-002"/><Con from="12" weight="-1.93973166920975e-003"/><Con from="13" weight="-1.94313708446391e-002"/><Con from="14" weight="-1.16454659160076e-002"/><Con from="15" weight="-6.31135349018358e-003"/><Con from="16" weight="-6.54177202583997e-003"/><Con from="17" weight="-2.30212786213320e-002"/><Con from="18" weight="-2.32481805412589e-002"/><Con from="19" weight="-2.05182414496624e-001"/><Con from="20" weight="-3.21631284721934e-002"/><Con from="21" weight="8.31102574156874e-002"/><Con from="22" weight="8.99828500644422e-002"/><Con from="23" weight="-2.74864265146564e-002"/><Con from="24" weight="9.08907859990291e-002"/><Con from="25" weight="1.50046805070629e-002"/><Con from="26" weight="1.46146710163955e-002"/><Con from="27" weight="2.08641567911763e-002"/><Con from="28" weight="4.75094818590004e-002"/><Con from="29" weight="2.19410209579044e-001"/><Con from="30" weight="-3.49104885994334e-002"/><Con from="31" weight="-1.61897528312940e-001"/><Con from="32" weight="6.66612961971092e-002"/><Con from="33" weight="1.33867426185386e-002"/><Con from="34" weight="-4.95126644632670e-003"/><Con from="35" weight="1.44241098736300e-001"/><Con from="36" weight="-5.80765147399657e-003"/><Con from="37" weight="-1.23552053409324e-001"/><Con from="38" weight="-1.10013575440638e-002"/><Con from="39" weight="9.47053912625876e-003"/><Con from="40" weight="-6.46271088517869e-002"/><Con from="41" weight="-2.17685277178425e-002"/><Con from="42" weight="-5.00276326236441e-002"/><Con from="43" weight="-5.41449884065013e-002"/><Con from="44" weight="-2.86946348642998e-002"/><Con from="45" weight="-2.77954836830792e-002"/><Con from="46" weight="5.49289354437446e-002"/><Con from="47" weight="-7.36942368345373e-002"/><Con from="48" weight="7.25415188755036e-002"/><Con from="49" weight="7.39532092646837e-002"/><Con from="50" weight="8.66965772087907e-002"/><Con from="51" weight="-6.71932397485942e-002"/><Con from="52" weight="5.82004156198659e-002"/><Con from="53" weight="1.16244566996887e-002"/><Con from="54" weight="8.05848183798510e-002"/><Con from="55" weight="-5.38257770606553e-002"/><Con from="56" weight="-6.58544203636648e-002"/><Con from="57" weight="-5.13012665746998e-002"/><Con from="58" weight="9.44070633641476e-002"/><Con from="59" weight="-1.94718691725788e-001"/><Con from="60" weight="3.83250844184021e-002"/><Con from="61" weight="-8.77188554203475e-003"/><Con from="62" weight="-

7.99034151603698e-002"/><Con from="63" weight="1.08786299828846e-001"/><Con
from="64" weight="1.28172083154600e-001"/><Con from="65" weight="-
9.43017790394798e-002"/><Con from="66" weight="-1.55382711747306e-002"/><Con
from="67" weight="7.33484462653279e-003"/><Con from="68" weight="-
7.78502293337303e-003"/><Con from="69" weight="-2.97635488151776e-002"/><Con
from="70" weight="4.40118638266806e-002"/><Con from="71"
weight="3.70709448606820e-002"/><Con from="72" weight="-2.13852849724401e-
001"/><Con from="73" weight="-5.82612145675993e-003"/><Con from="74"
weight="1.76434874824129e-002"/><Con from="75" weight="1.42563646233308e-
002"/><Con from="76" weight="1.99359173452040e-003"/><Con from="77"
weight="1.99019746287219e-002"/><Con from="78" weight="-9.75556937674598e-
003"/><Con from="79" weight="2.11667178231324e-002"/><Con from="80" weight="-
2.10929128221588e-002"/><Con from="81" weight="6.34290814627172e-002"/><Con
from="82" weight="-1.09860969430441e-002"/><Con from="83"
weight="2.20642267327912e-002"/><Con from="84" weight="-2.25047696459839e-
002"/><Con from="85" weight="-8.45263044218439e-003"/><Con from="86"
weight="1.64181183897210e-002"/><Con from="87" weight="7.08605932197443e-
002"/><Con from="88" weight="1.17831989332472e-002"/><Con from="89"
weight="1.84099557420930e-002"/><Con from="90" weight="7.65264731944470e-
004"/><Con from="91" weight="-4.31911735050344e-002"/><Con from="92" weight="-
6.17913840148582e-003"/><Con from="93" weight="-4.41247972331958e-002"/><Con
from="94" weight="4.17882821910164e-003"/><Con from="95"
weight="3.46858558647025e-003"/><Con from="96" weight="-2.37527585317128e-
002"/><Con from="97" weight="-3.47451535587402e-002"/><Con from="98"
weight="8.52976899759139e-003"/><Con from="99" weight="-6.32738822950009e-
002"/><Con from="100" weight="-5.96946085797823e-002"/><Con from="101"
weight="-2.21100061117248e-002"/><Con from="102" weight="5.11644066744281e-
004"/><Con from="103" weight="2.71572800728307e-002"/><Con from="104"
weight="-1.88952020528399e-002"/><Con from="105" weight="6.86101505491833e-
002"/><Con from="106" weight="-9.39902762590530e-004"/><Con from="107"
weight="-1.43245281716976e-002"/><Con from="108" weight="8.95306635293537e-
003"/><Con from="109" weight="3.56893202183190e-002"/><Con from="110"
weight="4.97253498075074e-002"/><Con from="111" weight="1.18147222826878e-
002"/><Con from="112" weight="5.81991928974660e-002"/><Con from="113"
weight="4.28653673411579e-003"/><Con from="114" weight="-2.79500859883013e-
004"/><Con from="115" weight="1.49616773717684e-002"/><Con from="116"
weight="-2.66689940647048e-002"/><Con from="117" weight="-9.91343240020798e-
002"/><Con from="118" weight="-1.67755818417534e-002"/><Con from="119"
weight="-3.76256163955821e-002"/><Con from="120" weight="6.34747935443396e-
003"/><Con from="121" weight="8.99895879166921e-002"/><Con from="122"
weight="-7.20306160908717e-003"/><Con from="123" weight="7.44873393624691e-
003"/><Con from="124" weight="1.77123538848872e-002"/><Con from="125"
weight="5.16256740949181e-002"/><Con from="126" weight="5.76519489556127e-
003"/><Con from="127" weight="2.48381051120855e-002"/><Con from="128"
weight="-8.75201900295808e-003"/><Con from="129" weight="-2.50589046738915e-
002"/><Con from="130" weight="4.59897282788517e-003"/><Con from="131"
weight="-1.13055020210858e-002"/><Con from="132" weight="2.15064269502501e-
002"/><Con from="133" weight="1.51804945933846e-003"/><Con from="134"
weight="3.10724212112272e-002"/><Con from="135" weight="2.80142295191433e-
002"/><Con from="136" weight="1.34074438599992e-002"/><Con from="137"
weight="-6.69521519903077e-002"/><Con from="138" weight="-5.64240511766740e-
002"/><Con from="139" weight="2.59298228609554e-002"/><Con from="140"
weight="-2.52707192717449e-003"/><Con from="141" weight="3.18175137883996e-
002"/><Con from="142" weight="-3.31826865436569e-002"/><Con from="143"
weight="-7.85030103970404e-004"/><Con from="144" weight="1.21346781911843e-
001"/><Con from="145" weight="9.46816741404161e-003"/><Con from="146"
weight="-1.09276021890429e-002"/><Con from="147" weight="-1.13056997827943e-
001"/><Con from="148" weight="-1.38263978605872e-002"/><Con from="149"
weight="-2.62867448901460e-002"/><Con from="150" weight="3.19657603995214e-
002"/><Con from="151" weight="1.34506254720455e-002"/><Con from="152"
weight="7.99376304122491e-003"/><Con from="153" weight="-4.79498575632528e-
002"/><Con from="154" weight="7.16426745226380e-004"/><Con from="155"
weight="2.78078255094445e-002"/><Con from="156" weight="-5.16600200053082e-
002"/><Con from="157" weight="-7.72064604108101e-003"/><Con from="158"
weight="3.78872213090811e-002"/><Con from="159" weight="2.11578857550569e-

002"/><Con from="160" weight="2.85160258120556e-003"/><Con from="161"
weight="1.00601290143586e-002"/><Con from="162" weight="-1.62115649560198e-
003"/><Con from="163" weight="-9.60949298498871e-002"/><Con from="164"
weight="1.75589303744918e-003"/><Con from="165" weight="2.06404663897010e-
002"/><Con from="166" weight="-3.42669606554395e-003"/><Con from="167"
weight="3.68702645725003e-002"/><Con from="168" weight="-6.57484817143552e-
003"/><Con from="169" weight="-4.74585131422107e-002"/><Con from="170"
weight="-2.04123094607483e-002"/><Con from="171" weight="-7.80066463681036e-
002"/><Con from="172" weight="-2.73921527846439e-003"/><Con from="173"
weight="3.95147604322807e-002"/><Con from="174" weight="-9.88793188350052e-
003"/><Con from="175" weight="1.09442330914577e-002"/><Con from="176"
weight="3.03789067095411e-003"/><Con from="177" weight="3.02289671273783e-
002"/></Neuron><Neuron id="181" bias="-9.00182350030797e-002"><Con from="0"
weight="-7.57782951795886e-002"/><Con from="1" weight="-4.06996889673397e-
002"/><Con from="2" weight="-6.75927589732570e-002"/><Con from="3" weight="-
2.62859185721317e-002"/><Con from="4" weight="-2.73977831094717e-002"/><Con
from="5" weight="-2.46041105191588e-002"/><Con from="6" weight="-
4.26277144065795e-002"/><Con from="7" weight="-1.99304403874394e-002"/><Con
from="8" weight="-2.70285514033355e-002"/><Con from="9" weight="-
9.53148226515640e-002"/><Con from="10" weight="-7.76006429985736e-002"/><Con
from="11" weight="-9.85952318005472e-003"/><Con from="12" weight="-
5.13482566800309e-003"/><Con from="13" weight="3.62337084556680e-003"/><Con
from="14" weight="8.19915049477222e-003"/><Con from="15"
weight="1.01009910195562e-002"/><Con from="16" weight="-1.83091486629086e-
003"/><Con from="17" weight="-9.83762762721405e-003"/><Con from="18" weight="-
1.19177743551552e-002"/><Con from="19" weight="-3.65668609283481e-002"/><Con
from="20" weight="-2.46323619932949e-002"/><Con from="21"
weight="3.46140831991955e-002"/><Con from="22" weight="-1.04915386088743e-
002"/><Con from="23" weight="-8.77909188732765e-002"/><Con from="24"
weight="3.79784304792767e-002"/><Con from="25" weight="6.88613404123895e-
003"/><Con from="26" weight="1.44183801149175e-002"/><Con from="27"
weight="1.79746434233390e-002"/><Con from="28" weight="1.41047361227510e-
002"/><Con from="29" weight="3.94861361416654e-002"/><Con from="30" weight="-
1.23762454806253e-002"/><Con from="31" weight="-5.65925569571772e-002"/><Con
from="32" weight="2.06900849015256e-002"/><Con from="33"
weight="7.96887245455953e-003"/><Con from="34" weight="-1.70348408314181e-
003"/><Con from="35" weight="4.46271090578053e-002"/><Con from="36" weight="-
7.23050541551670e-003"/><Con from="37" weight="-5.03081134385746e-002"/><Con
from="38" weight="-5.89974804078093e-002"/><Con from="39"
weight="6.09595614323282e-004"/><Con from="40" weight="-2.53656243510715e-
002"/><Con from="41" weight="-1.40027047552313e-002"/><Con from="42" weight="-
2.69899589191865e-002"/><Con from="43" weight="-4.52450997539226e-002"/><Con
from="44" weight="-1.45987941820161e-002"/><Con from="45" weight="-
2.63667854420746e-003"/><Con from="46" weight="-3.84593184966760e-003"/><Con
from="47" weight="-3.94632489006638e-002"/><Con from="48"
weight="1.86129846286985e-002"/><Con from="49" weight="1.11415944137087e-
002"/><Con from="50" weight="-3.60233818157351e-002"/><Con from="51" weight="-
3.47854965774177e-002"/><Con from="52" weight="2.35001779662249e-002"/><Con
from="53" weight="-4.34966596992687e-004"/><Con from="54"
weight="3.94490186393943e-003"/><Con from="55" weight="-7.19018104126293e-
003"/><Con from="56" weight="-3.06411064938522e-002"/><Con from="57" weight="-
6.44993167867088e-003"/><Con from="58" weight="5.76484588775166e-002"/><Con
from="59" weight="-2.92664282683360e-002"/><Con from="60"
weight="3.93859574457379e-003"/><Con from="61" weight="-2.24820454577286e-
002"/><Con from="62" weight="-3.01981413504372e-002"/><Con from="63" weight="-
9.16684658548551e-002"/><Con from="64" weight="3.83727785536976e-004"/><Con
from="65" weight="-3.88564575551901e-002"/><Con from="66" weight="-
1.53022291540413e-002"/><Con from="67" weight="-1.45864398003331e-003"/><Con
from="68" weight="-1.66115328567350e-002"/><Con from="69" weight="-
2.63902100247669e-002"/><Con from="70" weight="1.00588726344974e-002"/><Con
from="71" weight="4.22320354203309e-003"/><Con from="72" weight="-
3.80148344927089e-002"/><Con from="73" weight="-2.26877976794811e-002"/><Con
from="74" weight="6.19953275245919e-003"/><Con from="75"
weight="1.03952671926853e-002"/><Con from="76" weight="-3.93140299414866e-
003"/><Con from="77" weight="1.02881981655288e-002"/><Con from="78" weight="-

1.71284441228943e-002"/><Con from="79" weight="1.90089639025259e-002"/><Con
from="80" weight="-1.00176446612334e-002"/><Con from="81"
weight="2.48987329221972e-002"/><Con from="82" weight="-4.98465849619623e-
003"/><Con from="83" weight="1.08826702955720e-002"/><Con from="84" weight="-
6.81155425088939e-003"/><Con from="85" weight="5.64940549494032e-003"/><Con
from="86" weight="1.80917208959914e-002"/><Con from="87" weight="-
3.61110340950204e-002"/><Con from="88" weight="8.39114325028323e-003"/><Con
from="89" weight="1.61299178410785e-003"/><Con from="90"
weight="1.47493015026605e-003"/><Con from="91" weight="-1.68747246319291e-
002"/><Con from="92" weight="-6.89271560403246e-003"/><Con from="93" weight="-
2.96637092928598e-002"/><Con from="94" weight="1.34537356424888e-002"/><Con
from="95" weight="-9.18999705809797e-003"/><Con from="96"
weight="9.33297422095709e-003"/><Con from="97" weight="-1.68886015705026e-
002"/><Con from="98" weight="5.30987601875340e-003"/><Con from="99" weight="-
3.83899278969171e-002"/><Con from="100" weight="-3.25207165812375e-002"/><Con
from="101" weight="-1.05120192122347e-003"/><Con from="102"
weight="2.38635812553119e-003"/><Con from="103" weight="1.10309801966631e-
002"/><Con from="104" weight="-2.70803108239921e-002"/><Con from="105"
weight="-8.13825056732057e-003"/><Con from="106" weight="7.211826186121 87e-
003"/><Con from="107" weight="-9.39275863073063e-003"/><Con from="108"
weight="3.96518036723994e-003"/><Con from="109" weight="8.68477012149536e-
003"/><Con from="110" weight="1.08735365132839e-002"/><Con from="111"
weight="-5.87747944801839e-003"/><Con from="112" weight="1.70208351700512e-
002"/><Con from="113" weight="-6.42714001436804e-003"/><Con from="114"
weight="-1.74969364817104e-003"/><Con from="115" weight="-3.83671549908846e-
003"/><Con from="116" weight="-4.18274362467647e-003"/><Con from="117"
weight="-3.91618319607287e-002"/><Con from="118" weight="-2.09664851707326e-
003"/><Con from="119" weight="-1.52569659949586e-002"/><Con from="120"
weight="-9.53996351115806e-003"/><Con from="121" weight="4.61569833865994e-
002"/><Con from="122" weight="-7.06149446114009e-003"/><Con from="123"
weight="-4.35459945676972e-003"/><Con from="124" weight="-3.97978002648001e-
003"/><Con from="125" weight="5.19044014269940e-003"/><Con from="126"
weight="8.00972376114830e-003"/><Con from="127" weight="7.50776766450535e-
003"/><Con from="128" weight="2.50807551834078e-003"/><Con from="129"
weight="-8.97430529962575e-003"/><Con from="130" weight="-1.86248224783557e-
003"/><Con from="131" weight="-3.18107974152399e-003"/><Con from="132"
weight="1.06344730472274e-002"/><Con from="133" weight="4.29850373737534e-
003"/><Con from="134" weight="4.45274152648736e-005"/><Con from="135"
weight="6.77688892500972e-003"/><Con from="136" weight="1.07378503061359e-
002"/><Con from="137" weight="-1.26591958947477e-002"/><Con from="138"
weight="-2.16024796375166e-002"/><Con from="139" weight="3.23115526516199e-
002"/><Con from="140" weight="8.73530546425722e-003"/><Con from="141"
weight="9.92585576853261e-003"/><Con from="142" weight="-1.78213234919922e-
002"/><Con from="143" weight="2.51181350118527e-004"/><Con from="144"
weight="1.44336514274772e-002"/><Con from="145" weight="8.85862326758849e-
003"/><Con from="146" weight="-7.61675109987598e-004"/><Con from="147"
weight="-5.31100566142868e-002"/><Con from="148" weight="-3.50926817445060e-
003"/><Con from="149" weight="-3.25969323345453e-003"/><Con from="150"
weight="6.86845996196040e-003"/><Con from="151" weight="-2.53972513916480e-
003"/><Con from="152" weight="-2.59476715060809e-003"/><Con from="153"
weight="-1.85174282783522e-002"/><Con from="154" weight="6.88423287868525e-
003"/><Con from="155" weight="8.76864868635195e-003"/><Con from="156"
weight="-1.28879694010551e-002"/><Con from="157" weight="-3.65190549783970e-
003"/><Con from="158" weight="3.12654608776001e-003"/><Con from="159"
weight="-2.28542074557388e-003"/><Con from="160" weight="1.11887949012521e-
002"/><Con from="161" weight="5.09245058499177e-004"/><Con from="162"
weight="-2.47162287676869e-003"/><Con from="163" weight="-8.16591879193904e-
003"/><Con from="164" weight="-1.03367138466455e-002"/><Con from="165"
weight="1.23116993356004e-002"/><Con from="166" weight="3.44539397096174e-
003"/><Con from="167" weight="1.19701372475245e-002"/><Con from="168"
weight="1.77539910683032e-002"/><Con from="169" weight="-2.02884776202454e-
002"/><Con from="170" weight="-1.71817181420163e-002"/><Con from="171"
weight="-2.49971594825612e-002"/><Con from="172" weight="1.54096215812807e-
003"/><Con from="173" weight="1.29851433385818e-002"/><Con from="174"
weight="-8.83000694862907e-003"/><Con from="175" weight="1.43372715135042e-

002"/><Con from="176" weight="-7.95527634936553e-003"/><Con from="177"
weight="1.79651898855891e-003"/></Neuron><Neuron id="182"
bias="4.25631396910834e-004"><Con from="0" weight="-1.20270662336291e-
001"/><Con from="1" weight="8.79433743366200e-002"/><Con from="2" weight="-
1.14584669077203e-001"/><Con from="3" weight="1.01364024043786e-001"/><Con
from="4" weight="8.25142881406331e-002"/><Con from="5"
weight="2.55866319712064e-002"/><Con from="6" weight="3.39486395662155e-
002"/><Con from="7" weight="-1.78011617744835e-002"/><Con from="8"
weight="4.55927046877660e-002"/><Con from="9" weight="2.61881349157215e-
002"/><Con from="10" weight="-1.05305243531060e-001"/><Con from="11"
weight="1.81123188287119e-003"/><Con from="12" weight="1.17774382853696e-
002"/><Con from="13" weight="-3.51701552331363e-002"/><Con from="14"
weight="3.01085490201742e-003"/><Con from="15" weight="-8.66410587401034e-
003"/><Con from="16" weight="6.44495054211740e-003"/><Con from="17"
weight="1.47089730955521e-002"/><Con from="18" weight="-3.40903010517559e-
003"/><Con from="19" weight="-1.00922894240732e-001"/><Con from="20" weight="-
6.98271087072639e-003"/><Con from="21" weight="2.80240395774732e-002"/><Con
from="22" weight="8.41977326405686e-002"/><Con from="23" weight="-
3.55531346654100e-002"/><Con from="24" weight="3.32725553897595e-002"/><Con
from="25" weight="1.98178826507964e-002"/><Con from="26" weight="-
7.11732120640379e-004"/><Con from="27" weight="6.85550910800695e-003"/><Con
from="28" weight="2.12200161133802e-002"/><Con from="29"
weight="2.26412767434643e-001"/><Con from="30" weight="-1.55059093989333e-
003"/><Con from="31" weight="-3.03563938818473e-002"/><Con from="32"
weight="3.35674078689712e-002"/><Con from="33" weight="1.10694306780814e-
002"/><Con from="34" weight="-6.38224629190786e-003"/><Con from="35"
weight="8.66175138390749e-002"/><Con from="36" weight="8.59527076657269e-
004"/><Con from="37" weight="-4.36589665900037e-002"/><Con from="38" weight="-
2.28192854892620e-001"/><Con from="39" weight="1.03979426123982e-002"/><Con
from="40" weight="-1.52233497928968e-002"/><Con from="41"
weight="4.95212013326717e-003"/><Con from="42" weight="-1.77892699187565e-
002"/><Con from="43" weight="-6.05949171615320e-002"/><Con from="44" weight="-
1.14344621048137e-002"/><Con from="45" weight="-1.52672882670584e-003"/><Con
from="46" weight="2.95708666365542e-002"/><Con from="47" weight="-
7.14047062316870e-003"/><Con from="48" weight="-1.08212890484381e-002"/><Con
from="49" weight="9.27096560697060e-002"/><Con from="50" weight="-
1.66455564330128e-002"/><Con from="51" weight="-5.34101804006734e-002"/><Con
from="52" weight="3.37210537079052e-002"/><Con from="53" weight="-
2.09277326867056e-003"/><Con from="54" weight="7.44198423914100e-002"/><Con
from="55" weight="-2.89198645954392e-002"/><Con from="56" weight="-
1.90501274265478e-002"/><Con from="57" weight="1.90842607122400e-002"/><Con
from="58" weight="1.44586672357109e-001"/><Con from="59" weight="-
1.03165038486272e-001"/><Con from="60" weight="2.07763606130569e-002"/><Con
from="61" weight="6.83484585939284e-003"/><Con from="62" weight="-
4.13986741293905e-002"/><Con from="63" weight="-2.56748082157067e-002"/><Con
from="64" weight="5.98992186195722e-004"/><Con from="65" weight="-
5.28192436283960e-002"/><Con from="66" weight="1.13207334191943e-002"/><Con
from="67" weight="6.67581263732607e-003"/><Con from="68" weight="-
2.69279634233346e-003"/><Con from="69" weight="3.82475743809707e-005"/><Con
from="70" weight="1.80322898101241e-002"/><Con from="71"
weight="1.21517627229206e-002"/><Con from="72" weight="-9.59947844882107e-
002"/><Con from="73" weight="-3.17065655804133e-002"/><Con from="74"
weight="2.69143515316373e-002"/><Con from="75" weight="2.77108684321888e-
003"/><Con from="76" weight="-5.75733180912950e-003"/><Con from="77" weight="-
1.54077261969154e-002"/><Con from="78" weight="2.09750585847738e-003"/><Con
from="79" weight="7.53582387215830e-002"/><Con from="80" weight="-
1.50266278802470e-002"/><Con from="81" weight="5.25600261675623e-002"/><Con
from="82" weight="2.21282643059870e-002"/><Con from="83"
weight="5.44602195123571e-003"/><Con from="84" weight="-4.39460824434543e-
003"/><Con from="85" weight="-1.64538715637822e-002"/><Con from="86" weight="-
4.59438967890675e-003"/><Con from="87" weight="4.26166345739221e-002"/><Con
from="88" weight="1.10372234011702e-002"/><Con from="89"
weight="1.49928062779381e-002"/><Con from="90" weight="9.38230075208245e-
003"/><Con from="91" weight="-1.28231374880826e-002"/><Con from="92" weight="-
3.36822763652592e-004"/><Con from="93" weight="-2.76977128021750e-002"/><Con

from="94" weight="-2.19608888714398e-002"/><Con from="95"
weight="2.92851766249547e-003"/><Con from="96" weight="1.09722040232478e-
002"/><Con from="97" weight="-2.27856654140576e-002"/><Con from="98" weight="-
1.72370457644224e-002"/><Con from="99" weight="-4.89684335253459e-002"/><Con
from="100" weight="-3.31950526010079e-002"/><Con from="101"
weight="4.75154028936443e-003"/><Con from="102" weight="2.81442594301831e-
003"/><Con from="103" weight="-7.23468045979000e-004"/><Con from="104"
weight="-7.78041546811975e-002"/><Con from="105" weight="8.05905566549009e-
002"/><Con from="106" weight="3.26689635145806e-002"/><Con from="107"
weight="-1.12274466299822e-002"/><Con from="108" weight="4.01032078160351e-
003"/><Con from="109" weight="2.11703524563341e-002"/><Con from="110"
weight="4.34142896748578e-002"/><Con from="111" weight="-5.37843748827696e-
003"/><Con from="112" weight="2.11604746760816e-002"/><Con from="113"
weight="-2.25085323170760e-002"/><Con from="114" weight="-1.01181582583898e-
002"/><Con from="115" weight="-1.57211597863094e-003"/><Con from="116"
weight="8.58608278479948e-003"/><Con from="117" weight="-1.80226993321658e-
002"/><Con from="118" weight="7.50081065833809e-003"/><Con from="119"
weight="-1.29914845255810e-002"/><Con from="120" weight="-6.37467969532257e-
003"/><Con from="121" weight="6.15870951474661e-002"/><Con from="122"
weight="-1.66657005131066e-002"/><Con from="123" weight="-2.09939794689712e-
003"/><Con from="124" weight="-5.08212204095134e-003"/><Con from="125"
weight="9.30133930090399e-003"/><Con from="126" weight="-2.26241810118191e-
002"/><Con from="127" weight="2.76253900281196e-003"/><Con from="128"
weight="-1.82640779305151e-003"/><Con from="129" weight="-4.48018680673489e-
003"/><Con from="130" weight="-5.81825316663769e-004"/><Con from="131"
weight="9.56115746181410e-004"/><Con from="132" weight="6.09400210614242e-
003"/><Con from="133" weight="2.28341317022791e-003"/><Con from="134"
weight="5.77043358541308e-003"/><Con from="135" weight="6.18319464929848e-
003"/><Con from="136" weight="1.49967305814443e-002"/><Con from="137"
weight="-2.30230110321888e-002"/><Con from="138" weight="-1.67493095696264e-
002"/><Con from="139" weight="1.96414511099411e-002"/><Con from="140"
weight="1.05293083965476e-002"/><Con from="141" weight="-7.20203854492568e-
003"/><Con from="142" weight="-2.48191201029858e-002"/><Con from="143"
weight="4.16417149900211e-004"/><Con from="144" weight="1.82504172176288e-
002"/><Con from="145" weight="-8.39925548891511e-003"/><Con from="146"
weight="-1.04237900924260e-002"/><Con from="147" weight="-8.09417645769423e-
002"/><Con from="148" weight="-7.72486952458145e-003"/><Con from="149"
weight="-4.27413007057179e-003"/><Con from="150" weight="1.90828961124700e-
003"/><Con from="151" weight="2.34372723806421e-002"/><Con from="152"
weight="-1.02354881953248e-002"/><Con from="153" weight="1.31034745520901e-
003"/><Con from="154" weight="1.18108307293387e-002"/><Con from="155"
weight="8.48239834620966e-003"/><Con from="156" weight="-1.91112410760193e-
002"/><Con from="157" weight="6.23277041165202e-003"/><Con from="158"
weight="1.38558570305361e-002"/><Con from="159" weight="1.34167585025593e-
002"/><Con from="160" weight="5.95149353978088e-003"/><Con from="161"
weight="1.49894072208384e-002"/><Con from="162" weight="1.27561751392423e-
002"/><Con from="163" weight="-2.30953318141693e-002"/><Con from="164"
weight="7.34686446097897e-003"/><Con from="165" weight="-4.55404667824951e-
004"/><Con from="166" weight="3.46554793972172e-003"/><Con from="167"
weight="-5.59328522392242e-003"/><Con from="168" weight="-7.55286684508573e-
003"/><Con from="169" weight="1.65154509834968e-003"/><Con from="170"
weight="2.25984311901111e-003"/><Con from="171" weight="-5.70357918603963e-
002"/><Con from="172" weight="1.09965286550466e-002"/><Con from="173"
weight="3.24954356831833e-003"/><Con from="174" weight="-3.99469168900173e-
003"/><Con from="175" weight="3.78509112154637e-003"/><Con from="176"
weight="8.53446864810746e-003"/><Con from="177" weight="9.01524739160396e-
003"/></Neuron><Neuron id="183" bias="1.55268088561858e-001"><Con from="0"
weight="3.44923741632942e-001"/><Con from="1" weight="2.92061296606508e-
002"/><Con from="2" weight="3.50835598156660e-001"/><Con from="3" weight="-
2.68614205584708e-003"/><Con from="4" weight="-4.46986420064543e-003"/><Con
from="5" weight="-4.51798140338392e-002"/><Con from="6" weight="-
3.22688644012983e-003"/><Con from="7" weight="6.40126663127280e-002"/><Con
from="8" weight="-6.88647317540161e-002"/><Con from="9"
weight="1.52147266901448e-001"/><Con from="10" weight="2.55577917448703e-
001"/><Con from="11" weight="5.56129598627692e-002"/><Con from="12"

weight="1.30832514175948e-002"/><Con from="13" weight="4.40861531899191e-002"/><Con from="14" weight="2.72742001736954e-002"/><Con from="15" weight="-4.17466285989853e-003"/><Con from="16" weight="-5.82001224949857e-003"/><Con from="17" weight="3.89930946695115e-002"/><Con from="18" weight="7.92442598878725e-002"/><Con from="19" weight="3.70993934746477e-001"/><Con from="20" weight="4.80088768630935e-002"/><Con from="21" weight="-1.72606442291776e-001"/><Con from="22" weight="-1.17882682826390e-001"/><Con from="23" weight="1.02576192683866e-001"/><Con from="24" weight="-1.23362729475337e-001"/><Con from="25" weight="-1.61656727749726e-002"/><Con from="26" weight="-3.48838240827700e-002"/><Con from="27" weight="-3.44962305771730e-002"/><Con from="28" weight="-1.08211451368528e-001"/><Con from="29" weight="-4.06295769991693e-001"/><Con from="30" weight="5.95980774233642e-002"/><Con from="31" weight="3.05168626575803e-001"/><Con from="32" weight="-1.28527149245700e-001"/><Con from="33" weight="-5.58037661408042e-002"/><Con from="34" weight="1.01681073627680e-003"/><Con from="35" weight="-2.99697355368946e-001"/><Con from="36" weight="-2.64782359771714e-003"/><Con from="37" weight="2.09548523630723e-001"/><Con from="38" weight="2.04707590362551e-001"/><Con from="39" weight="1.41622951116985e-003"/><Con from="40" weight="1.14002478491503e-001"/><Con from="41" weight="5.29778694347053e-002"/><Con from="42" weight="6.69393504077331e-002"/><Con from="43" weight="1.26355505173594e-001"/><Con from="44" weight="6.51271953239781e-002"/><Con from="45" weight="5.02749669768164e-002"/><Con from="46" weight="-4.90146403911204e-002"/><Con from="47" weight="9.93149803416223e-002"/><Con from="48" weight="-8.80833887462396e-002"/><Con from="49" weight="-6.69031852805055e-002"/><Con from="50" weight="-7.34906762438510e-002"/><Con from="51" weight="1.44179469761647e-001"/><Con from="52" weight="-9.53115638274542e-002"/><Con from="53" weight="-3.17992472413973e-002"/><Con from="54" weight="-1.45192031022136e-001"/><Con from="55" weight="1.06647009744965e-001"/><Con from="56" weight="1.54821237200894e-001"/><Con from="57" weight="9.11829937711720e-002"/><Con from="58" weight="-6.22283403113892e-002"/><Con from="59" weight="3.19859455350981e-001"/><Con from="60" weight="-1.07510150942405e-001"/><Con from="61" weight="2.11080546432706e-002"/><Con from="62" weight="1.28245878400365e-001"/><Con from="63" weight="-1.69235517958434e-001"/><Con from="64" weight="-1.58486163134190e-001"/><Con from="65" weight="1.49330962619368e-001"/><Con from="66" weight="1.31710571425827e-002"/><Con from="67" weight="-2.89382553601708e-003"/><Con from="68" weight="4.90464311703781e-003"/><Con from="69" weight="6.56600284000595e-002"/><Con from="70" weight="-1.04070434617061e-001"/><Con from="71" weight="-2.97375748047734e-002"/><Con from="72" weight="3.60325229064107e-001"/><Con from="73" weight="5.22013830936140e-002"/><Con from="74" weight="4.43186159990131e-002"/><Con from="75" weight="-1.02253061448757e-002"/><Con from="76" weight="2.15236462287332e-002"/><Con from="77" weight="3.74698573792064e-003"/><Con from="78" weight="3.82713352132871e-002"/><Con from="79" weight="1.33542918950883e-002"/><Con from="80" weight="4.78440521971868e-002"/><Con from="81" weight="-1.35909433925763e-001"/><Con from="82" weight="8.80314840369968e-005"/><Con from="83" weight="-2.24501050669552e-002"/><Con from="84" weight="2.29685602403989e-002"/><Con from="85" weight="9.73058266465645e-003"/><Con from="86" weight="-1.13629209912061e-002"/><Con from="87" weight="-1.49741709386691e-001"/><Con from="88" weight="-2.76892787206060e-002"/><Con from="89" weight="-1.41932425660012e-002"/><Con from="90" weight="-2.68597772257244e-003"/><Con from="91" weight="8.06940471551711e-002"/><Con from="92" weight="-7.02984201825074e-003"/><Con from="93" weight="8.34138934165284e-002"/><Con from="94" weight="-7.00804451342710e-003"/><Con from="95" weight="4.49731175921690e-003"/><Con from="96" weight="2.90181747739188e-002"/><Con from="97" weight="5.11232347681209e-002"/><Con from="98" weight="2.16442002152510e-003"/><Con from="99" weight="1.30774478449734e-001"/><Con from="100" weight="9.43471679523164e-002"/><Con from="101" weight="7.14333106099929e-003"/><Con from="102" weight="-5.80204130427081e-003"/><Con from="103" weight="-3.06647100888762e-002"/><Con from="104" weight="4.69982055118197e-002"/><Con from="105" weight="-8.16397589437015e-002"/><Con from="106" weight="-2.58088917954013e-002"/><Con from="107" weight="2.88195130426694e-002"/><Con from="108" weight="-1.70061605589989e-002"/><Con from="109" weight="-5.72574729763919e-

002"/><Con from="110" weight="-5.72736399919161e-002"/><Con from="111"
weight="4.00630582904587e-003"/><Con from="112" weight="-9.31336479381579e-
002"/><Con from="113" weight="-2.19158281594048e-003"/><Con from="114"
weight="-1.50778598770054e-002"/><Con from="115" weight="-2.78773070002402e-
002"/><Con from="116" weight="5.08958724356245e-002"/><Con from="117"
weight="1.54784014514854e-001"/><Con from="118" weight="-4.07711462116599e-
004"/><Con from="119" weight="8.11023103484506e-002"/><Con from="120"
weight="7.30572029757068e-003"/><Con from="121" weight="-1.50791349969006e-
001"/><Con from="122" weight="4.60093848241981e-002"/><Con from="123"
weight="5.62491404845643e-003"/><Con from="124" weight="-7.52259153001606e-
003"/><Con from="125" weight="-1.19937601905760e-001"/><Con from="126"
weight="8.06078405605264e-003"/><Con from="127" weight="-2.13060408816312e-
002"/><Con from="128" weight="-2.04423167189114e-002"/><Con from="129"
weight="5.92157429568847e-002"/><Con from="130" weight="5.47931124286207e-
003"/><Con from="131" weight="1.28230276289018e-002"/><Con from="132"
weight="-3.03609018190813e-002"/><Con from="133" weight="5.33813918591972e-
003"/><Con from="134" weight="-6.77970609700523e-002"/><Con from="135"
weight="-6.96280155874042e-002"/><Con from="136" weight="-3.41855354060772e-
002"/><Con from="137" weight="9.95165208891653e-002"/><Con from="138"
weight="1.17930466665282e-001"/><Con from="139" weight="-7.35751054715012e-
002"/><Con from="140" weight="3.29945146674344e-004"/><Con from="141"
weight="-3.58853511737301e-002"/><Con from="142" weight="6.76852814473072e-
002"/><Con from="143" weight="2.99638223042097e-004"/><Con from="144"
weight="-2.01724236712374e-001"/><Con from="145" weight="8.48988377229745e-
003"/><Con from="146" weight="5.22221548214029e-003"/><Con from="147"
weight="1.88706216002479e-001"/><Con from="148" weight="3.01551554044158e-
002"/><Con from="149" weight="4.86122641846937e-002"/><Con from="150"
weight="-5.96328617069886e-002"/><Con from="151" weight="-9.16616710928116e-
003"/><Con from="152" weight="7.95377528904682e-003"/><Con from="153"
weight="7.03328172064913e-002"/><Con from="154" weight="-3.30837185277683e-
002"/><Con from="155" weight="-3.07608101129588e-002"/><Con from="156"
weight="8.51611352020543e-002"/><Con from="157" weight="2.27250239774470e-
003"/><Con from="158" weight="-8.47371803158945e-002"/><Con from="159"
weight="-3.46311907556725e-002"/><Con from="160" weight="-4.43200187685418e-
002"/><Con from="161" weight="-3.45512610566601e-002"/><Con from="162"
weight="1.26660780714243e-002"/><Con from="163" weight="1.63748174613021e-
001"/><Con from="164" weight="-2.24636352973999e-002"/><Con from="165"
weight="-2.20385811972945e-002"/><Con from="166" weight="-2.44359031059631e-
003"/><Con from="167" weight="-6.34453594163322e-002"/><Con from="168"
weight="-5.00719518719448e-003"/><Con from="169" weight="6.51994605211254e-
002"/><Con from="170" weight="2.61902717301609e-002"/><Con from="171"
weight="1.34209858075231e-001"/><Con from="172" weight="7.35907723964156e-
003"/><Con from="173" weight="-5.41008564499016e-002"/><Con from="174"
weight="1.10383776812220e-002"/><Con from="175" weight="-5.06178353698227e-
002"/><Con from="176" weight="-1.66505275942887e-002"/><Con from="177"
weight="-3.05012775880191e-002"/></Neuron><Neuron id="184" bias="-
2.75498740321410e-001"><Con from="0" weight="-3.63396443740842e-001"/><Con
from="1" weight="2.32543749077720e-002"/><Con from="2" weight="-
3.93542468041094e-001"/><Con from="3" weight="9.28509990340607e-002"/><Con
from="4" weight="5.21017695389319e-002"/><Con from="5" weight="-
9.85590896072279e-002"/><Con from="6" weight="-5.41251191776473e-002"/><Con
from="7" weight="-1.16391476554019e-001"/><Con from="8" weight="-
6.56260500536317e-002"/><Con from="9" weight="-2.46423764059576e-001"/><Con
from="10" weight="-3.43158351519610e-001"/><Con from="11" weight="-
1.93445189099240e-002"/><Con from="12" weight="8.83680909074903e-003"/><Con
from="13" weight="-4.95214446454748e-002"/><Con from="14"
weight="7.43719291420655e-003"/><Con from="15" weight="1.59985860987671e-
002"/><Con from="16" weight="-1.07668856120524e-002"/><Con from="17"
weight="3.13208062990107e-003"/><Con from="18" weight="1.58729937225519e-
003"/><Con from="19" weight="-2.40545780640174e-001"/><Con from="20" weight="-
1.68044535492049e-002"/><Con from="21" weight="8.69542881481993e-002"/><Con
from="22" weight="9.78154025865865e-002"/><Con from="23" weight="-
2.80024682116246e-001"/><Con from="24" weight="9.39949122506300e-002"/><Con
from="25" weight="3.20349569261626e-003"/><Con from="26"
weight="2.54162587832258e-002"/><Con from="27" weight="4.11248390861611e-

003"/><Con from="28" weight="2.49485550308233e-002"/><Con from="29"
weight="4.51697740961976e-001"/><Con from="30" weight="-2.13798341491046e-
002"/><Con from="31" weight="-1.17770980157351e-001"/><Con from="32"
weight="6.00739499638322e-002"/><Con from="33" weight="1.53342475095549e-
002"/><Con from="34" weight="-1.69343058433205e-003"/><Con from="35"
weight="1.93792384704639e-001"/><Con from="36" weight="-1.61615003978261e-
003"/><Con from="37" weight="-1.20866590210620e-001"/><Con from="38" weight="-
6.19149274835326e-001"/><Con from="39" weight="-4.02074449236795e-004"/><Con
from="40" weight="-7.99927166373016e-002"/><Con from="41" weight="-
4.23652455366092e-003"/><Con from="42" weight="-4.37126687486590e-002"/><Con
from="43" weight="-1.98500407305592e-001"/><Con from="44" weight="-
7.49848576272806e-002"/><Con from="45" weight="-2.00477563968770e-002"/><Con
from="46" weight="1.86330626515729e-002"/><Con from="47" weight="-
6.87957658193903e-002"/><Con from="48" weight="-2.80958770184362e-002"/><Con
from="49" weight="1.29904939136497e-001"/><Con from="50" weight="-
1.57042537886737e-001"/><Con from="51" weight="-1.47770681180439e-001"/><Con
from="52" weight="3.51008975234580e-002"/><Con from="53"
weight="2.91320689735550e-002"/><Con from="54" weight="1.29626387877221e-
001"/><Con from="55" weight="-5.06279155046103e-002"/><Con from="56" weight="-
7.73082776251051e-002"/><Con from="57" weight="1.04857212070012e-002"/><Con
from="58" weight="3.18503774081550e-001"/><Con from="59" weight="-
2.14298518801751e-001"/><Con from="60" weight="4.73050156293131e-002"/><Con
from="61" weight="-3.24323103004023e-002"/><Con from="62" weight="-
8.18428125247755e-002"/><Con from="63" weight="-2.92831897254523e-001"/><Con
from="64" weight="-2.82802027120745e-002"/><Con from="65" weight="-
1.46159638746319e-001"/><Con from="66" weight="7.75995135578307e-003"/><Con
from="67" weight="4.06003863463304e-004"/><Con from="68" weight="-
1.18786030463745e-002"/><Con from="69" weight="-4.30827347857982e-002"/><Con
from="70" weight="5.18896124736722e-002"/><Con from="71"
weight="1.98675171937676e-002"/><Con from="72" weight="-2.28592465703452e-
001"/><Con from="73" weight="-1.35717629860723e-001"/><Con from="74"
weight="3.88376497309334e-002"/><Con from="75" weight="1.62550826404162e-
002"/><Con from="76" weight="-2.82572114518653e-002"/><Con from="77" weight="-
4.91824944245619e-004"/><Con from="78" weight="2.10585694365401e-003"/><Con
from="79" weight="1.68209045938620e-001"/><Con from="80" weight="-
3.57849336945967e-002"/><Con from="81" weight="1.28534301346499e-001"/><Con
from="82" weight="3.00924311613737e-002"/><Con from="83"
weight="1.59573589600372e-002"/><Con from="84" weight="-2.34836578204876e-
002"/><Con from="85" weight="-7.91468506239782e-003"/><Con from="86"
weight="2.02397062170909e-002"/><Con from="87" weight="-3.56050939125071e-
003"/><Con from="88" weight="-1.25093942865562e-003"/><Con from="89"
weight="2.65772993287323e-003"/><Con from="90" weight="2.14600755397482e-
002"/><Con from="91" weight="-3.49834441284265e-002"/><Con from="92" weight="-
7.92565518469328e-003"/><Con from="93" weight="-3.75890370156133e-002"/><Con
from="94" weight="4.00759312982198e-003"/><Con from="95" weight="-
1.87596984081356e-003"/><Con from="96" weight="-1.56860581511929e-002"/><Con
from="97" weight="-2.65153861822248e-002"/><Con from="98" weight="-
1.94303683213251e-002"/><Con from="99" weight="-1.11488361944168e-001"/><Con
from="100" weight="-8.11204472662080e-002"/><Con from="101" weight="-
6.42504972606229e-004"/><Con from="102" weight="-5.66961493203835e-003"/><Con
from="103" weight="1.55042111892219e-002"/><Con from="104" weight="-
2.20179504919313e-001"/><Con from="105" weight="7.49285729753377e-002"/><Con
from="106" weight="8.61731029395893e-002"/><Con from="107" weight="-
1.20795497689034e-002"/><Con from="108" weight="6.05895589661144e-003"/><Con
from="109" weight="2.34905798955673e-002"/><Con from="110"
weight="8.10491042067941e-002"/><Con from="111" weight="-1.11565998896933e-
002"/><Con from="112" weight="4.86879135101824e-002"/><Con from="113"
weight="-1.19367604469691e-002"/><Con from="114" weight="-3.83456570822181e-
003"/><Con from="115" weight="-9.09968860345717e-003"/><Con from="116"
weight="-1.65653807742407e-002"/><Con from="117" weight="-6.54667247998130e-
002"/><Con from="118" weight="5.40955131825776e-003"/><Con from="119"
weight="-3.12463595156105e-002"/><Con from="120" weight="2.18193651002735e-
003"/><Con from="121" weight="1.31353239468955e-001"/><Con from="122"
weight="-8.22777849981742e-002"/><Con from="123" weight="-6.91150945543904e-
003"/><Con from="124" weight="-7.64532483689800e-003"/><Con from="125"

weight="4.78578520409839e-002"/><Con from="126" weight="-1.66817680759122e-003"/><Con from="127" weight="-9.54643501493965e-003"/><Con from="128" weight="-4.28216767888121e-003"/><Con from="129" weight="-4.33350714339302e-002"/><Con from="130" weight="-3.47893104352357e-003"/><Con from="131" weight="3.79919496112922e-003"/><Con from="132" weight="3.26856106766792e-002"/><Con from="133" weight="-1.09127825489423e-002"/><Con from="134" weight="1.54118148145402e-002"/><Con from="135" weight="3.18368856983273e-002"/><Con from="136" weight="1.82471810664792e-002"/><Con from="137" weight="-4.93681673686499e-002"/><Con from="138" weight="-5.11395289042305e-002"/><Con from="139" weight="5.46585808938365e-002"/><Con from="140" weight="-1.33261227091972e-002"/><Con from="141" weight="3.77843878667355e-003"/><Con from="142" weight="-5.33757900902911e-002"/><Con from="143" weight="-3.43011260810371e-003"/><Con from="144" weight="3.36603394287119e-002"/><Con from="145" weight="-1.84680196796393e-003"/><Con from="146" weight="3.51456183202058e-003"/><Con from="147" weight="-1.97529081178110e-001"/><Con from="148" weight="-3.68675544111522e-004"/><Con from="149" weight="-3.06931162344691e-002"/><Con from="150" weight="2.96155566599739e-002"/><Con from="151" weight="4.65036196708987e-002"/><Con from="152" weight="-2.66948418683939e-002"/><Con from="153" weight="-2.56817698853459e-002"/><Con from="154" weight="-6.04538101889174e-004"/><Con from="155" weight="-5.97922783437616e-003"/><Con from="156" weight="-4.28160332118420e-002"/><Con from="157" weight="1.07632074994717e-002"/><Con from="158" weight="3.53889938854501e-002"/><Con from="159" weight="7.43654587219064e-003"/><Con from="160" weight="1.73247258591557e-002"/><Con from="161" weight="1.71140292039365e-002"/><Con from="162" weight="5.46255857279075e-003"/><Con from="163" weight="-6.52081019196970e-002"/><Con from="164" weight="1.83850221592296e-002"/><Con from="165" weight="1.42446371197149e-002"/><Con from="166" weight="5.46204282789151e-003"/><Con from="167" weight="7.14503067517812e-003"/><Con from="168" weight="-1.90043190175181e-004"/><Con from="169" weight="-4.11169930989208e-002"/><Con from="170" weight="-1.42920910357595e-002"/><Con from="171" weight="-1.21787325319122e-001"/><Con from="172" weight="-1.01285122502947e-003"/><Con from="173" weight="2.51486533828891e-002"/><Con from="174" weight="-1.27598893827714e-002"/><Con from="175" weight="3.41309262006146e-002"/><Con from="176" weight="-3.69220598792872e-004"/><Con from="177" weight="5.87170581509960e-002"/></Neuron></NeuralLayer><NeuralLayer numberOfNeurons="1" activationFunction="sine"><Neuron id="185" bias="2.45984517062421e+000"><Con from="178" weight="-4.40871660857188e-002"/><Con from="179" weight="-1.77405126458416e-001"/><Con from="180" weight="-2.44801672949688e-001"/><Con from="181" weight="-1.95788812824293e-001"/><Con from="182" weight="9.62822274264705e-002"/><Con from="183" weight="2.26937368807419e-001"/><Con from="184" weight="7.37445845409614e-002"/></Neuron></NeuralLayer><NeuralOutputs numberOfOutputs="1"><NeuralOutput outputNeuron="185"><DerivedField optype="continuous"><NormContinuous field="Bulk_density" shift="-5.04273504273504e-001" scale="8.54700854700855e-001"><LinearNorm orig="5.90000000000000e-001" norm="0.00000000000000e+000"/><LinearNorm orig="1.76000000000000e+000" norm="1.00000000000000e+000"/></NormContinuous></DerivedField></NeuralOutput></NeuralOutputs></NeuralNetwork></PMML>

A.2 Random Forest - R Script

The basic code used to generate the Random Forest Bulk Density predictions is shown below. From this basic script all Random Forest Models were developed, using different datasets and input variables relevant to the particular model.

```
library(randomForest)
which(sapply(Topsoil, function(y) nlevels(y) > 32))
train = Topsoil[ c(1:239), ]
test = Topsoil[ c(240:342), ]
set.seed(100)
bulk.rf<-randomForest(Bulk_density ~ RCS + AAR + AT0_ANNUAL + FCD_MED + + PSMD + PT + Soil_association + Great_group + land_use + Aspect + Curvature + Iwahashi + Pennock + Slope + STI + Elevation + PM1, data=train, ntree=1000, mtry=2, importance=TRUE, proximity=TRUE, varUsed=TRUE, varImpPlot=TRUE)
print(bulk.rf)
varImpPlot(bulk.rf)
A_predict = predict(bulk.rf, test)
A_predict
```

A.3 Multiple Linear Regression - R Script

The basic code used to generate the Multiple Linear Regression Bulk Density predictions is shown below. From this basic script all MLR Models were developed, using different datasets and input variables relevant to the particular model.

```
set.seed(100)
fit<-lm(Bulk_density ~ factor(LU_GROUP) + factor(Great_group) + AAR + AT0_ANNUAL + FCD_MED + PSMD + PT + Curvature + Profile + Slope + STI + SWI + Elevation, data=train)
summary(fit)
layout(matrix(c(1,2,3,4),2,2))
plot(fit)
#stepwise variable selection
```

```
library(MASS)
step <- stepAIC(fit, direction="both")
step$anova
```

Appendix B - Chapter 3

B.1 Conditional Probability Tables for the Optimised Naive BN

The conditional probability tables for the optimised Bayesian Network are shown below.

Table B.1-1: CPT for the 'LEX' British Geological Survey rock lexicon node of the Optimised Naive BN (key available to download from http://www.bgs.ac.uk/lexicon/)

LEX	Bulk Density				
	0.59 to 0.99	0.99 to 1.14	1.14 to 1.28	1.28 to 1.41	1.41 to 1.76
AS	16.12	16.71	16.41	34.04	16.71
AW	21.56	11.17	21.95	34.14	11.17
BAN	36.75	12.69	24.94	12.93	12.69
BCMU	16.17	8.38	24.69	25.61	25.14
BLCR	13.92	21.65	35.43	7.35	21.65
BLL	27.85	28.86	14.17	14.70	14.43
BMS	4.17	8.65	12.75	48.47	25.96
BSG	54.67	11.33	11.13	11.54	11.33
CBRD	19.43	20.14	19.78	20.51	20.14
CDF	19.43	20.14	19.78	20.51	20.14
CHAM	19.47	10.09	29.72	20.55	20.17
CHG	19.43	20.14	19.78	20.51	20.14
CLT	13.89	28.79	28.27	14.66	14.39
CTM	19.43	20.14	19.78	20.51	20.14
DYS	36.75	12.69	24.94	12.93	12.69
ECL	19.43	20.14	19.78	20.51	20.14
EDW	13.85	14.35	28.20	29.24	14.35
EN	12.14	12.58	24.72	12.82	37.75
ETM	16.17	33.53	16.46	17.07	16.76
GUN	9.68	30.10	9.86	10.22	40.14
HA	13.85	14.36	14.10	14.62	43.07
HANS	16.12	16.71	16.41	34.04	16.71
HBR	32.54	16.86	16.56	17.17	16.86
KDM	27.85	28.86	14.17	14.70	14.43
KHS	12.11	25.11	12.33	12.79	37.66
LES	32.54	16.86	16.56	17.17	16.86
LLUS	12.11	25.11	24.66	25.57	12.55
LOS	32.54	16.86	16.56	17.17	16.86
MI	19.43	20.14	19.78	20.51	20.14
MMG	17.24	22.75	22.34	16.55	21.12

MO	12.14	37.75	24.72	12.82	12.58
MOI	16.12	16.71	16.41	34.04	16.71
MORRI	61.66	19.17	6.28	6.51	6.39
MRB	27.92	14.47	28.42	14.73	14.47
MVC	32.54	16.86	16.56	17.17	16.86
NS	13.78	14.28	14.03	43.63	14.28
NTC	8.04	16.66	8.18	25.46	41.66
ONS	32.54	16.86	16.56	17.17	16.86
OWSH	27.92	14.47	28.42	14.73	14.47
PET	16.22	16.81	33.03	17.12	16.81
PLCM	13.85	43.07	14.10	14.62	14.36
PLD	32.54	16.86	16.56	17.17	16.86
PLWF	19.43	20.14	19.78	20.51	20.14
PMCM	12.11	25.11	12.33	12.79	37.66
RG	12.14	25.16	24.72	12.82	25.16
RLS	12.09	25.05	12.30	25.51	25.05
RR	13.82	28.64	14.06	29.17	14.32
SASH	13.85	28.71	14.10	14.62	28.71
SIM	15.30	26.43	20.77	26.92	10.57
SMG	16.12	16.71	16.41	34.04	16.71
SPPS	16.22	16.81	33.03	17.12	16.81
TLM	19.43	20.14	19.78	20.51	20.14
TPSF	16.17	16.76	16.46	17.07	33.53
ULUS	19.43	20.14	19.78	20.51	20.14
WBY	16.25	33.68	33.08	8.58	8.42
WCT	27.77	14.39	14.13	29.31	14.39
WDF	54.67	11.33	11.13	11.54	11.33
WEL	19.43	20.14	19.78	20.51	20.14
WGF	13.85	14.35	28.20	29.24	14.35
WHM	26.66	27.63	27.13	9.38	9.21
WIT	13.89	14.39	28.27	14.66	28.79
WOL	16.22	16.81	33.03	17.12	16.81
WRS	6.04	18.77	18.43	25.49	31.28

Table B.1-2: CPT for the 'Soil Association' node of the Optimised Naive BN

Soil Association	Bulk Density				
	0.59 to 0.99	**0.99 to 1.14**	**1.14 to 1.28**	**1.28 to 1.41**	**1.41 to 1.76**
Cambic stagnogley soils	30.31	32.02	5.19	16.47	16.01
Cambic stagnohumic gley soils	65.43	8.64	8.40	8.89	8.64
Ferritic brown earths	23.98	12.66	24.63	26.06	12.66
Gleyic brown earths	19.13	20.21	19.66	20.79	20.21
Humo-ferric podzols	15.91	33.62	16.35	17.30	16.81
Ironpan stagnopodzols	32.12	16.96	16.50	17.46	16.96
Man made soils	15.91	16.81	16.35	17.30	33.62
Paleo-argillic stagnogley soils	13.62	14.39	27.99	29.61	14.39
Pelo-alluvial gley soils	61.20	12.93	6.29	6.65	12.93
Pelo-stagnogley soils	38.82	24.60	19.94	8.44	8.20
Stagnogleyic argillic brown earths	4.43	21.04	34.11	24.06	16.37
Typical argillic brown earths	13.57	14.33	13.94	29.50	28.66
Typical argillic pelosols	20.56	21.71	21.12	14.90	21.71
Typical brown alluvial soils	10.64	33.73	32.81	11.57	11.24
Typical brown calcareous earths	15.99	16.89	32.85	17.38	16.89
Typical brown earths	12.70	16.76	19.57	24.15	26.82
Typical brown podzolic soils	32.12	16.96	16.50	17.46	16.96
Typical brown sands	4.08	12.94	4.20	39.95	38.83
Typical calcareous pelosols	12.69	26.80	13.03	20.68	26.80
Typical cambic gley soils	15.84	16.73	16.27	34.43	16.73
Typical humic-sandy gley soils	13.62	43.18	14.00	14.81	14.39
Typical paleo-argillic brown earths	19.13	20.21	19.66	20.79	20.21
Typical sandy gley soils	13.57	14.33	13.94	29.50	28.66
Typical stagnogley soils	18.76	19.82	31.33	12.75	17.34

Table B.1-3: CPT for the 'Parent Material' node of the Optimised Naive BN (see Table B.1-4 for the key to the Parent Material Code)

Parent Material Code	Bulk Density				
	0.59 to 0.99	**0.99 to 1.14**	**1.14 to 1.28**	**1.28 to 1.41**	**1.41 to 1.76**
Bb	32.01	16.99	16.48	17.53	16.99
Bg	23.89	12.68	24.60	26.16	12.68
Bh	26.17	37.04	17.97	9.55	9.26
Bo	13.44	14.27	9.23	39.27	23.79
Bp	19.05	20.23	19.62	20.87	20.23
Cf	19.05	20.23	19.62	20.87	20.23
Da	15.93	16.91	32.81	17.45	16.91
Db	11.76	24.98	12.12	38.65	12.49
Ea	48.65	20.66	15.03	5.33	10.33
Ef	27.36	29.04	14.09	14.98	14.52
Eg	8.63	18.32	26.66	18.90	27.48
Ei	11.45	20.66	23.59	17.56	26.74
Fi	27.23	20.81	20.19	17.89	13.87
Fq	14.23	10.07	19.53	25.97	30.20
Fw	13.50	14.34	13.91	29.58	28.67
Fx	11.81	25.08	12.16	25.87	25.08
Fy	10.57	22.43	32.64	23.14	11.22
Ga	15.85	16.82	16.32	17.36	33.65

Table B.1-4: Key for Parent Material classes

Code	Parent Material
Bb	basic crystalline rock
Bg	ironstone
Bh	limestone
Bo	sandstone
Bp	siltstone
Cf	very hard siliceous stones
Cg	sandstones, siltstones, mudstones or slate
Da	calcareous gravel
Db	non-calcareous gravel
Ea	river alluvium
Ef	stoneless drift
Eg	chalky drift
Ei	drift with siliceous stones
Fi	clay or soft mudstone
Fq	sand or soft sandstone
Fw	soft siltstone or shale
Fx	soft siltstone and sandstone
Fy	soft shale or siltstone
Ga	replaced material

Table B.1-5: CPT for the 'Land Use' node of the Optimised Naive BN (see Table B.1-6 for the key to the Land Use Code)

Land Use Code	Bulk Density				
	0.59 to 0.99	**0.99 to 1.14**	**1.14 to 1.28**	**1.28 to 1.41**	**1.41 to 1.76**
AR	9.03	9.62	18.64	25.59	37.12
CO	31.93	17.01	16.47	17.59	17.01
DC	42.87	22.84	11.06	11.81	11.42
FA	6.25	19.99	12.90	27.55	33.31
GC	15.80	16.83	16.30	17.40	33.67
HC	15.71	16.74	16.21	34.61	16.74
LE	2.41	12.82	24.83	29.17	30.77
OR	15.80	16.83	16.30	17.40	33.67
OT	37.06	31.59	15.29	8.16	7.90
PG	30.26	28.56	22.30	14.29	4.61
RC	18.99	20.24	19.60	20.93	20.24
RG	41.30	14.67	14.20	15.17	14.67
T?	13.58	28.95	28.03	14.96	14.47
UG	31.93	17.01	16.47	17.59	17.01

Table B.1-6: Key for Land Use classes

Code	Land Use
AR	arable
CO	coniferous
DC	deciduous
FA	fallow
GC	green crops
HC	horticultural crops
LE	ley grassland
OR	orchard
OT	other
PG	permanent grassland
RC	root crops
RG	rough grazing
T?	other tillage
UG	upland grass

Table B.1-7: CPT for the 'SWI' Saga Wetness Index node of the Optimised Naive BN

SWI	Bulk Density				
	0.59 to 0.99	**0.99 to 1.14**	**1.14 to 1.28**	**1.28 to 1.41**	**1.41 to 1.76**
10.4 to 14.9	33.49	17.0	20.10	21.71	7.59
14.9 to 15.6	16.84	30.23	7.76	20.96	24.18
15.6 to 16	15.65	10.53	22.31	24.10	27.39
16 to 16.6	15.44	18.47	28.46	17.29	20.32
16.6 to 18.3	18.19	23.16	20.58	16.67	21.37

Table B.1-8: CPT for the 'FCD_MED' Annual median number of field capacity days node of the Optimised Naive BN

FCD_MED	Bulk Density				
	0.59 to 0.99	**0.99 to 1.14**	**1.14 to 1.28**	**1.28 to 1.41**	**1.41 to 1.76**
122 to 141	12.24	17.58	19.04	15.99	35.15
141 to 149	8.25	22.21	14.97	32.34	22.21
149 to 155	18.87	18.47	24.90	21.13	16.62
155 to 171	8.34	19.77	24.23	24.30	23.36
171 to 278	46.67	21.78	16.13	8.71	6.70

Table B.1-9: CPT for the 'Curvature' node of the Optimised Naive BN

Curvature	Bulk Density				
	0.59 to 0.99	**0.99 to 1.14**	**1.14 to 1.28**	**1.28 to 1.41**	**1.41 to 1.76**
-4 to -0.3	33.95	21.94	11.74	17.75	14.63
-0.3 to 0	22.75	22.46	25.56	16.99	12.25
0	13.42	14.45	27.83	25.04	19.26
0 to 0.3	20.43	26.40	6.36	16.02	30.80
0.3 to 4.7	14.92	17.67	23.20	21.72	22.49

Table B.1-10: CPT for the 'AAR' Average Annual Rainfall node of the Optimised Naive BN

AAR	Bulk Density				
	0.59 to 0.99	0.99 to 1.14	1.14 to 1.28	1.28 to 1.41	1.41 to 1.76
570 to 650	6.22	13.39	19.34	20.89	40.17
650 to 665	27.12	17.03	23.43	20.25	12.17
665 to 678	10.87	23.41	22.55	26.78	16.39
678 to 720	8.47	25.53	17.56	26.55	21.88
720 to 1270	45.21	21.10	18.76	8.44	6.49

Table B.1-11: CPT for the 'Elevation' node of the Optimised Naive BN

Elevation	Bulk Density				
	0.59 to 0.99	0.99 to 1.14	1.14 to 1.28	1.28 to 1.41	1.41 to 1.76
9 to 53	9.18	19.76	24.74	22.61	23.71
53 to 77	18.78	9.19	15.93	24.85	31.25
77 to 105	7.30	31.46	18.93	26.58	15.73
105 to 130	17.67	21.14	20.36	17.59	23.25
130 to 410	44.12	19.36	20.33	9.15	7.04

Table B.1-12: CPT for the 'Bulk Density' node of the Optimised Naive BN

Bulk Density				
0.59 to 0.99	0.99 to 1.14	1.14 to 1.28	1.28 to 1.41	1.41 to 1.76
21.31	19.67	20.49	18.85	19.67

B.2 Conditional Probability Table for the Expert Structure model

In the expert structured BN, only the 'bulk density' node relates directly to D_b hence it is the only CPT displayed here (Table B.1-1). The Bulk density node had two parent nodes (Land use and Soil association), however, the CPT associated with the node is much larger than those of the Naive Network (Appendix B.1). Other nodes in the Expert BN have many more parent nodes meaning the CPTs associated with those nodes are very large. As they do not relate to D_b they have been omitted from this Abstract.

Table B.2-1: CPT for the 'Bulk Density' node of the Expert structured BN

Soil Association	Land Use	Bulk Density				
		0.59 to 0.99	0.99 to 1.14	1.14 to 1.28	1.28 to 1.41	1.41 to 1.76
Cambic stagnogley soils	**AR**	0.00	2.00	10.87	22.73	64.41
Cambic stagnogley soils	**CO**	0.00	9.81	39.98	19.39	30.83
Cambic stagnogley soils	**DC**	0.00	9.81	39.98	19.39	30.83
Cambic stagnogley soils	**FA**	0.00	4.81	13.08	31.71	50.41
Cambic stagnogley soils	**GC**	0.00	4.94	13.42	19.53	62.11
Cambic stagnogley soils	**HC**	0.00	5.58	15.17	44.15	35.10
Cambic stagnogley soils	**LE**	0.00	2.56	16.21	30.33	50.90
Cambic stagnogley soils	**OR**	0.00	3.77	10.24	14.90	71.08
Cambic stagnogley soils	**OT**	0.00	21.16	28.76	27.90	22.18
Cambic stagnogley soils	**PG**	0.00	16.68	36.05	32.97	14.30
Cambic stagnogley soils	**RC**	0.00	3.42	18.58	13.52	64.48
Cambic stagnogley soils	**RG**	0.00	7.16	19.47	28.33	45.04
Cambic stagnogley soils	**T?**	0.00	16.06	29.10	21.17	33.66
Cambic stagnogley soils	**UG**	0.00	13.37	18.17	26.43	42.03
Cambic stagnohumic gley soils	**AR**	16.44	11.95	12.18	11.32	48.12
Cambic stagnohumic gley soils	**CO**	34.50	28.21	21.56	4.65	11.08
Cambic stagnohumic gley soils	**DC**	51.31	20.97	16.03	3.46	8.24
Cambic stagnohumic gley soils	**FA**	10.80	26.48	13.49	14.54	34.68
Cambic stagnohumic gley soils	**GC**	26.40	21.58	11.00	7.11	33.92
Cambic stagnohumic gley soils	**HC**	29.28	23.94	12.20	15.77	18.81
Cambic stagnohumic gley soils	**LE**	3.47	17.04	20.26	16.84	42.39
Cambic stagnohumic gley soils	**OR**	22.57	18.45	9.40	6.08	43.50

Cambic stagnohumic gley soils	**OT**	52.17	31.98	8.15	3.51	4.19
Cambic stagnohumic gley soils	**PG**	42.17	34.48	13.98	5.68	3.69
Cambic stagnohumic gley soils	**RC**	20.63	16.86	17.19	5.56	39.76
Cambic stagnohumic gley soils	**RG**	58.30	15.89	8.10	5.23	12.48
Cambic stagnohumic gley soils	**T?**	19.24	47.19	16.03	5.18	12.36
Cambic stagnohumic gley soils	**UG**	40.30	32.94	8.39	5.43	12.94
Ferritic brown earths	**AR**	13.03	18.93	32.16	35.87	0.00
Ferritic brown earths	**CO**	19.02	31.10	39.63	10.25	0.00
Ferritic brown earths	**DC**	31.97	26.13	33.29	8.61	0.00
Ferritic brown earths	**FA**	6.47	31.73	26.95	34.85	0.00
Ferritic brown earths	**GC**	19.60	32.05	27.22	21.12	0.00
Ferritic brown earths	**HC**	16.19	26.46	22.48	34.88	0.00
Ferritic brown earths	**LE**	2.01	19.76	39.16	39.07	0.00
Ferritic brown earths	**OR**	19.60	32.05	27.22	21.12	0.00
Ferritic brown earths	**OT**	33.15	40.65	17.26	8.93	0.00
Ferritic brown earths	**PG**	23.37	38.21	25.82	12.59	0.00
Ferritic brown earths	**RC**	15.41	25.19	42.79	16.60	0.00
Ferritic brown earths	**RG**	42.25	23.02	19.56	15.17	0.00
Ferritic brown earths	**T?**	10.25	50.26	28.46	11.04	0.00
Ferritic brown earths	**UG**	25.85	42.27	17.95	13.93	0.00
Gleyic brown earths	**AR**	13.03	18.93	32.16	35.87	0.00
Gleyic brown earths	**CO**	19.02	31.10	39.63	10.25	0.00
Gleyic brown earths	**DC**	31.97	26.13	33.29	8.61	0.00
Gleyic brown earths	**FA**	6.47	31.73	26.95	34.85	0.00
Gleyic brown earths	**GC**	19.60	32.05	27.22	21.12	0.00
Gleyic brown earths	**HC**	16.19	26.46	22.48	34.88	0.00
Gleyic brown earths	**LE**	2.01	19.76	39.16	39.07	0.00
Gleyic brown earths	**OR**	19.60	32.05	27.22	21.12	0.00
Gleyic brown earths	**OT**	33.15	40.65	17.26	8.93	0.00
Gleyic brown earths	**PG**	23.37	38.21	25.82	12.59	0.00
Gleyic brown earths	**RC**	15.41	25.19	42.79	16.60	0.00
Gleyic brown earths	**RG**	42.25	23.02	19.56	15.17	0.00
Gleyic brown earths	**T?**	10.25	50.26	28.46	11.04	0.00
Gleyic brown earths	**UG**	25.85	42.27	17.95	13.93	0.00
Humo-ferric podzols	**AR**	13.03	18.93	32.16	35.87	0.00
Humo-ferric podzols	**CO**	19.02	31.10	39.63	10.25	0.00
Humo-ferric podzols	**DC**	31.97	26.13	33.29	8.61	0.00
Humo-ferric podzols	**FA**	6.47	31.73	26.95	34.85	0.00
Humo-ferric podzols	**GC**	19.60	32.05	27.22	21.12	0.00
Humo-ferric podzols	**HC**	16.19	26.46	22.48	34.88	0.00
Humo-ferric podzols	**LE**	2.01	19.76	39.16	39.07	0.00
Humo-ferric podzols	**OR**	19.60	32.05	27.22	21.12	0.00
Humo-ferric podzols	**OT**	33.15	40.65	17.26	8.93	0.00
Humo-ferric podzols	**PG**	23.37	38.21	25.82	12.59	0.00

Humo-ferric podzols	**RC**	15.41	25.19	42.79	16.60	0.00
Humo-ferric podzols	**RG**	42.25	23.02	19.56	15.17	0.00
Humo-ferric podzols	**T?**	10.25	50.26	28.46	11.04	0.00
Humo-ferric podzols	**UG**	25.85	42.27	17.95	13.93	0.00
Ironpan stagnopodzols	**AR**	27.90	25.34	27.55	19.21	0.00
Ironpan stagnopodzols	**CO**	33.45	34.18	27.87	4.50	0.00
Ironpan stagnopodzols	**DC**	50.13	25.61	20.88	3.38	0.00
Ironpan stagnopodzols	**FA**	14.13	43.31	23.54	19.03	0.00
Ironpan stagnopodzols	**GC**	35.13	35.90	19.51	9.46	0.00
Ironpan stagnopodzols	**HC**	32.09	32.79	17.83	17.29	0.00
Ironpan stagnopodzols	**LE**	5.06	31.03	39.36	24.54	0.00
Ironpan stagnopodzols	**OR**	35.13	35.90	19.51	9.46	0.00
Ironpan stagnopodzols	**OT**	48.97	37.53	10.20	3.30	0.00
Ironpan stagnopodzols	**PG**	38.49	39.33	17.00	5.18	0.00
Ironpan stagnopodzols	**RC**	29.39	30.04	32.65	7.92	0.00
Ironpan stagnopodzols	**RG**	61.90	21.08	11.46	5.56	0.00
Ironpan stagnopodzols	**T?**	18.36	56.29	20.40	4.95	0.00
Ironpan stagnopodzols	**UG**	41.08	41.98	11.41	5.53	0.00
Man made soils	**AR**	7.72	2.81	5.72	15.95	67.80
Man made soils	**CO**	29.40	12.02	18.37	11.88	28.33
Man made soils	**DC**	45.44	9.29	14.20	9.18	21.89
Man made soils	**FA**	5.83	7.15	7.29	23.56	56.18
Man made soils	**GC**	15.41	6.30	6.42	12.46	59.41
Man made soils	**HC**	18.63	7.61	7.76	30.10	35.90
Man made soils	**LE**	1.65	4.06	9.65	24.06	60.58
Man made soils	**OR**	11.88	4.86	4.95	9.60	68.71
Man made soils	**OT**	52.47	16.09	8.20	10.60	12.64
Man made soils	**PG**	41.55	16.98	13.77	16.79	10.92
Man made soils	**RC**	11.32	4.63	9.43	9.15	65.47
Man made soils	**RG**	45.73	6.23	6.35	12.32	29.38
Man made soils	**T?**	17.26	21.16	14.38	13.95	33.26
Man made soils	**UG**	33.50	13.69	6.98	13.54	32.29
Paleo-argillic stagnogley soils	**AR**	0.00	2.00	10.87	22.73	64.41
Paleo-argillic stagnogley soils	**CO**	0.00	9.81	39.98	19.39	30.83
Paleo-argillic stagnogley soils	**DC**	0.00	9.81	39.98	19.39	30.83
Paleo-argillic stagnogley soils	**FA**	0.00	4.81	13.08	31.71	50.41
Paleo-argillic stagnogley soils	**GC**	0.00	4.94	13.42	19.53	62.11
Paleo-argillic stagnogley soils	**HC**	0.00	5.58	15.17	44.15	35.10
Paleo-argillic stagnogley soils	**LE**	0.00	2.56	16.21	30.33	50.90
Paleo-argillic stagnogley soils	**OR**	0.00	3.77	10.24	14.90	71.08
Paleo-argillic stagnogley soils	**OT**	0.00	21.16	28.76	27.90	22.18
Paleo-argillic stagnogley soils	**PG**	0.00	16.68	36.05	32.97	14.30
Paleo-argillic stagnogley soils	**RC**	0.00	3.42	18.58	13.52	64.48
Paleo-argillic stagnogley soils	**RG**	0.00	7.16	19.47	28.33	45.04

Paleo-argillic stagnogley soils	**T?**	0.00	16.06	29.10	21.17	33.66
Paleo-argillic stagnogley soils	**UG**	0.00	13.37	18.17	26.43	42.03
Pelo-alluvial gley soils	**AR**	53.85	19.56	26.59	0.00	0.00
Pelo-alluvial gley soils	**CO**	54.79	22.39	22.82	0.00	0.00
Pelo-alluvial gley soils	**DC**	70.79	14.47	14.74	0.00	0.00
Pelo-alluvial gley soils	**FA**	32.69	40.08	27.23	0.00	0.00
Pelo-alluvial gley soils	**GC**	59.30	24.24	16.47	0.00	0.00
Pelo-alluvial gley soils	**HC**	59.30	24.24	16.47	0.00	0.00
Pelo-alluvial gley soils	**LE**	13.62	33.41	52.97	0.00	0.00
Pelo-alluvial gley soils	**OR**	59.30	24.24	16.47	0.00	0.00
Pelo-alluvial gley soils	**OT**	70.89	21.73	7.38	0.00	0.00
Pelo-alluvial gley soils	**PG**	61.36	25.08	13.56	0.00	0.00
Pelo-alluvial gley soils	**RC**	50.91	20.81	28.28	0.00	0.00
Pelo-alluvial gley soils	**RG**	81.38	11.09	7.53	0.00	0.00
Pelo-alluvial gley soils	**T?**	35.95	44.08	19.97	0.00	0.00
Pelo-alluvial gley soils	**UG**	64.62	26.41	8.97	0.00	0.00
Pelo-stagnogley soils	**AR**	0.00	2.00	10.87	22.73	64.41
Pelo-stagnogley soils	**CO**	0.00	9.81	39.98	19.39	30.83
Pelo-stagnogley soils	**DC**	0.00	9.81	39.98	19.39	30.83
Pelo-stagnogley soils	**FA**	0.00	4.81	13.08	31.71	50.41
Pelo-stagnogley soils	**GC**	0.00	4.94	13.42	19.53	62.11
Pelo-stagnogley soils	**HC**	0.00	5.58	15.17	44.15	35.10
Pelo-stagnogley soils	**LE**	0.00	2.56	16.21	30.33	50.90
Pelo-stagnogley soils	**OR**	0.00	3.77	10.24	14.90	71.08
Pelo-stagnogley soils	**OT**	0.00	21.16	28.76	27.90	22.18
Pelo-stagnogley soils	**PG**	0.00	16.68	36.05	32.97	14.30
Pelo-stagnogley soils	**RC**	0.00	3.42	18.58	13.52	64.48
Pelo-stagnogley soils	**RG**	0.00	7.16	19.47	28.33	45.04
Pelo-stagnogley soils	**T?**	0.00	16.06	29.10	21.17	33.66
Pelo-stagnogley soils	**UG**	0.00	13.37	18.17	26.43	42.03
Stagnogleyic argillic brown earths	**AR**	0.00	5.51	18.73	31.34	44.41
Stagnogleyic argillic brown earths	**CO**	0.00	18.79	47.87	18.57	14.77
Stagnogleyic argillic brown earths	**DC**	0.00	18.79	47.87	18.57	14.77
Stagnogleyic argillic brown earths	**FA**	0.00	11.61	19.72	38.26	30.41
Stagnogleyic argillic brown earths	**GC**	0.00	12.79	21.72	25.29	40.20
Stagnogleyic argillic brown earths	**HC**	0.00	12.16	20.65	48.08	19.11
Stagnogleyic argillic brown earths	**LE**	0.00	6.30	24.97	37.37	31.36
Stagnogleyic argillic brown earths	**OR**	0.00	10.65	18.09	21.05	50.21
Stagnogleyic argillic brown earths	**OT**	0.00	36.10	30.66	23.79	9.46
Stagnogleyic argillic brown earths	**PG**	0.00	28.14	38.02	27.81	6.03
Stagnogleyic argillic brown earths	**RC**	0.00	9.02	30.64	17.83	42.52
Stagnogleyic argillic brown earths	**RG**	0.00	16.01	27.19	31.65	25.16
Stagnogleyic argillic brown earths	**T?**	0.00	30.16	34.16	19.88	15.80
Stagnogleyic argillic brown earths	**UG**	0.00	27.60	23.44	27.28	21.69

Typical argillic brown earths	**AR**	16.44	11.95	12.18	11.32	48.12
Typical argillic brown earths	**CO**	34.50	28.21	21.56	4.65	11.08
Typical argillic brown earths	**DC**	51.31	20.97	16.03	3.46	8.24
Typical argillic brown earths	**FA**	10.80	26.48	13.49	14.54	34.68
Typical argillic brown earths	**GC**	26.40	21.58	11.00	7.11	33.92
Typical argillic brown earths	**HC**	29.28	23.94	12.20	15.77	18.81
Typical argillic brown earths	**LE**	3.47	17.04	20.26	16.84	42.39
Typical argillic brown earths	**OR**	22.57	18.45	9.40	6.08	43.50
Typical argillic brown earths	**OT**	52.17	31.98	8.15	3.51	4.19
Typical argillic brown earths	**PG**	42.17	34.48	13.98	5.68	3.69
Typical argillic brown earths	**RC**	20.63	16.86	17.19	5.56	39.76
Typical argillic brown earths	**RG**	58.30	15.89	8.10	5.23	12.48
Typical argillic brown earths	**T?**	19.24	47.19	16.03	5.18	12.36
Typical argillic brown earths	**UG**	40.30	32.94	8.39	5.43	12.94
Typical argillic pelosols	**AR**	0.00	1.74	9.44	32.89	55.93
Typical argillic pelosols	**CO**	0.00	8.68	35.40	28.62	27.30
Typical argillic pelosols	**DC**	0.00	8.68	35.40	28.62	27.30
Typical argillic pelosols	**FA**	0.00	3.97	10.79	43.62	41.61
Typical argillic pelosols	**GC**	0.00	4.37	11.88	28.80	54.95
Typical argillic pelosols	**HC**	0.00	4.31	11.72	56.85	27.12
Typical argillic pelosols	**LE**	0.00	2.13	13.49	42.05	42.34
Typical argillic pelosols	**OR**	0.00	3.43	9.32	22.59	64.66
Typical argillic pelosols	**OT**	0.00	17.85	24.25	39.20	18.70
Typical argillic pelosols	**PG**	0.00	13.67	29.56	45.05	11.72
Typical argillic pelosols	**RC**	0.00	3.14	17.05	20.67	59.15
Typical argillic pelosols	**RG**	0.00	6.03	16.38	39.71	37.88
Typical argillic pelosols	**T?**	0.00	14.08	25.50	30.92	29.50
Typical argillic pelosols	**UG**	0.00	11.37	15.45	37.46	35.73
Typical brown alluvial soils	**AR**	53.85	19.56	26.59	0.00	0.00
Typical brown alluvial soils	**CO**	54.79	22.39	22.82	0.00	0.00
Typical brown alluvial soils	**DC**	70.79	14.47	14.74	0.00	0.00
Typical brown alluvial soils	**FA**	32.69	40.08	27.23	0.00	0.00
Typical brown alluvial soils	**GC**	59.30	24.24	16.47	0.00	0.00
Typical brown alluvial soils	**HC**	59.30	24.24	16.47	0.00	0.00
Typical brown alluvial soils	**LE**	13.62	33.41	52.97	0.00	0.00
Typical brown alluvial soils	**OR**	59.30	24.24	16.47	0.00	0.00
Typical brown alluvial soils	**OT**	70.89	21.73	7.38	0.00	0.00
Typical brown alluvial soils	**PG**	61.36	25.08	13.56	0.00	0.00
Typical brown alluvial soils	**RC**	50.91	20.81	28.28	0.00	0.00
Typical brown alluvial soils	**RG**	81.38	11.09	7.53	0.00	0.00
Typical brown alluvial soils	**T?**	35.95	44.08	19.97	0.00	0.00
Typical brown alluvial soils	**UG**	64.62	26.41	8.97	0.00	0.00
Typical brown calcareous earths	**AR**	0.00	13.32	36.20	50.48	0.00
Typical brown calcareous earths	**CO**	0.00	27.05	55.13	17.82	0.00

Typical brown calcareous earths	**DC**	0.00	27.05	55.13	17.82	0.00
Typical brown calcareous earths	**FA**	0.00	21.95	29.83	48.22	0.00
Typical brown calcareous earths	**GC**	0.00	27.20	36.96	35.85	0.00
Typical brown calcareous earths	**HC**	0.00	20.02	27.21	52.77	0.00
Typical brown calcareous earths	**LE**	0.00	12.31	39.03	48.67	0.00
Typical brown calcareous earths	**OR**	0.00	27.20	36.96	35.85	0.00
Typical brown calcareous earths	**OT**	0.00	47.20	32.07	20.74	0.00
Typical brown calcareous earths	**PG**	0.00	36.50	39.45	24.05	0.00
Typical brown calcareous earths	**RC**	0.00	19.86	53.97	26.17	0.00
Typical brown calcareous earths	**RG**	0.00	27.20	36.96	35.85	0.00
Typical brown calcareous earths	**T?**	0.00	42.64	38.63	18.73	0.00
Typical brown calcareous earths	**UG**	0.00	42.76	29.06	28.18	0.00
Typical brown earths	**AR**	13.03	18.93	32.16	35.87	0.00
Typical brown earths	**CO**	19.02	31.10	39.63	10.25	0.00
Typical brown earths	**DC**	31.97	26.13	33.29	8.61	0.00
Typical brown earths	**FA**	6.47	31.73	26.95	34.85	0.00
Typical brown earths	**GC**	19.60	32.05	27.22	21.12	0.00
Typical brown earths	**HC**	16.19	26.46	22.48	34.88	0.00
Typical brown earths	**LE**	2.01	19.76	39.16	39.07	0.00
Typical brown earths	**OR**	19.60	32.05	27.22	21.12	0.00
Typical brown earths	**OT**	33.15	40.65	17.26	8.93	0.00
Typical brown earths	**PG**	23.37	38.21	25.82	12.59	0.00
Typical brown earths	**RC**	15.41	25.19	42.79	16.60	0.00
Typical brown earths	**RG**	42.25	23.02	19.56	15.17	0.00
Typical brown earths	**T?**	10.25	50.26	28.46	11.04	0.00
Typical brown earths	**UG**	25.85	42.27	17.95	13.93	0.00
Typical brown podzolic soils	**AR**	13.68	14.91	33.76	37.66	0.00
Typical brown podzolic soils	**CO**	20.63	25.29	42.97	11.11	0.00
Typical brown podzolic soils	**DC**	34.20	20.97	35.62	9.21	0.00
Typical brown podzolic soils	**FA**	7.03	25.85	29.27	37.85	0.00
Typical brown podzolic soils	**GC**	21.31	26.13	29.59	22.96	0.00
Typical brown podzolic soils	**HC**	17.33	21.25	24.07	37.35	0.00
Typical brown podzolic soils	**LE**	2.12	15.59	41.19	41.10	0.00
Typical brown podzolic soils	**OR**	21.31	26.13	29.59	22.96	0.00
Typical brown podzolic soils	**OT**	36.90	33.94	19.22	9.94	0.00
Typical brown podzolic soils	**PG**	25.84	31.69	28.55	13.92	0.00
Typical brown podzolic soils	**RC**	16.45	20.16	45.67	17.72	0.00
Typical brown podzolic soils	**RG**	44.83	18.32	20.75	16.10	0.00
Typical brown podzolic soils	**T?**	11.72	43.11	32.55	12.63	0.00
Typical brown podzolic soils	**UG**	28.91	35.45	20.07	15.57	0.00
Typical brown sands	**AR**	16.44	11.95	12.18	11.32	48.12
Typical brown sands	**CO**	34.50	28.21	21.56	4.65	11.08
Typical brown sands	**DC**	51.31	20.97	16.03	3.46	8.24
Typical brown sands	**FA**	10.80	26.48	13.49	14.54	34.68

Typical brown sands	**GC**	26.40	21.58	11.00	7.11	33.92
Typical brown sands	**HC**	29.28	23.94	12.20	15.77	18.81
Typical brown sands	**LE**	3.47	17.04	20.26	16.84	42.39
Typical brown sands	**OR**	22.57	18.45	9.40	6.08	43.50
Typical brown sands	**OT**	52.17	31.98	8.15	3.51	4.19
Typical brown sands	**PG**	42.17	34.48	13.98	5.68	3.69
Typical brown sands	**RC**	20.63	16.86	17.19	5.56	39.76
Typical brown sands	**RG**	58.30	15.89	8.10	5.23	12.48
Typical brown sands	**T?**	19.24	47.19	16.03	5.18	12.36
Typical brown sands	**UG**	40.30	32.94	8.39	5.43	12.94
Typical calcareous pelosols	**AR**	0.00	1.86	10.10	28.17	59.87
Typical calcareous pelosols	**CO**	0.00	9.21	37.55	24.28	28.95
Typical calcareous pelosols	**DC**	0.00	9.21	37.55	24.28	28.95
Typical calcareous pelosols	**FA**	0.00	4.35	11.83	38.23	45.59
Typical calcareous pelosols	**GC**	0.00	4.64	12.60	24.45	58.31
Typical calcareous pelosols	**HC**	0.00	4.87	13.23	51.31	30.59
Typical calcareous pelosols	**LE**	0.00	2.32	14.73	36.73	46.23
Typical calcareous pelosols	**OR**	0.00	3.59	9.76	18.93	67.72
Typical calcareous pelosols	**OT**	0.00	19.36	26.31	34.03	20.29
Typical calcareous pelosols	**PG**	0.00	15.03	32.48	39.61	12.88
Typical calcareous pelosols	**RC**	0.00	3.27	17.78	17.25	61.70
Typical calcareous pelosols	**RG**	0.00	6.55	17.79	34.51	41.15
Typical calcareous pelosols	**T?**	0.00	15.00	27.19	26.37	31.44
Typical calcareous pelosols	**UG**	0.00	12.29	16.70	32.39	38.62
Typical cambic gley soils	**AR**	0.00	13.32	36.20	50.48	0.00
Typical cambic gley soils	**CO**	0.00	27.05	55.13	17.82	0.00
Typical cambic gley soils	**DC**	0.00	27.05	55.13	17.82	0.00
Typical cambic gley soils	**FA**	0.00	21.95	29.83	48.22	0.00
Typical cambic gley soils	**GC**	0.00	27.20	36.96	35.85	0.00
Typical cambic gley soils	**HC**	0.00	20.02	27.21	52.77	0.00
Typical cambic gley soils	**LE**	0.00	12.31	39.03	48.67	0.00
Typical cambic gley soils	**OR**	0.00	27.20	36.96	35.85	0.00
Typical cambic gley soils	**OT**	0.00	47.20	32.07	20.74	0.00
Typical cambic gley soils	**PG**	0.00	36.50	39.45	24.05	0.00
Typical cambic gley soils	**RC**	0.00	19.86	53.97	26.17	0.00
Typical cambic gley soils	**RG**	0.00	27.20	36.96	35.85	0.00
Typical cambic gley soils	**T?**	0.00	42.64	38.63	18.73	0.00
Typical cambic gley soils	**UG**	0.00	42.76	29.06	28.18	0.00
Typical humic-sandy gley soils	**AR**	53.85	19.56	26.59	0.00	0.00
Typical humic-sandy gley soils	**CO**	54.79	22.39	22.82	0.00	0.00
Typical humic-sandy gley soils	**DC**	70.79	14.47	14.74	0.00	0.00
Typical humic-sandy gley soils	**FA**	32.69	40.08	27.23	0.00	0.00
Typical humic-sandy gley soils	**GC**	59.30	24.24	16.47	0.00	0.00
Typical humic-sandy gley soils	**HC**	59.30	24.24	16.47	0.00	0.00

Typical humic-sandy gley soils	**LE**	13.62	33.41	52.97	0.00	0.00
Typical humic-sandy gley soils	**OR**	59.30	24.24	16.47	0.00	0.00
Typical humic-sandy gley soils	**OT**	70.89	21.73	7.38	0.00	0.00
Typical humic-sandy gley soils	**PG**	61.36	25.08	13.56	0.00	0.00
Typical humic-sandy gley soils	**RC**	50.91	20.81	28.28	0.00	0.00
Typical humic-sandy gley soils	**RG**	81.38	11.09	7.53	0.00	0.00
Typical humic-sandy gley soils	**T?**	35.95	44.08	19.97	0.00	0.00
Typical humic-sandy gley soils	**UG**	64.62	26.41	8.97	0.00	0.00
Typical paleo-argillic brown earths	**AR**	0.00	2.00	10.87	22.73	64.41
Typical paleo-argillic brown earths	**CO**	0.00	9.81	39.98	19.39	30.83
Typical paleo-argillic brown earths	**DC**	0.00	9.81	39.98	19.39	30.83
Typical paleo-argillic brown earths	**FA**	0.00	4.81	13.08	31.71	50.41
Typical paleo-argillic brown earths	**GC**	0.00	4.94	13.42	19.53	62.11
Typical paleo-argillic brown earths	**HC**	0.00	5.58	15.17	44.15	35.10
Typical paleo-argillic brown earths	**LE**	0.00	2.56	16.21	30.33	50.90
Typical paleo-argillic brown earths	**OR**	0.00	3.77	10.24	14.90	71.08
Typical paleo-argillic brown earths	**OT**	0.00	21.16	28.76	27.90	22.18
Typical paleo-argillic brown earths	**PG**	0.00	16.68	36.05	32.97	14.30
Typical paleo-argillic brown earths	**RC**	0.00	3.42	18.58	13.52	64.48
Typical paleo-argillic brown earths	**RG**	0.00	7.16	19.47	28.33	45.04
Typical paleo-argillic brown earths	**T?**	0.00	16.06	29.10	21.17	33.66
Typical paleo-argillic brown earths	**UG**	0.00	13.37	18.17	26.43	42.03
Typical sandy gley soils	**AR**	27.90	25.34	27.55	19.21	0.00
Typical sandy gley soils	**CO**	33.45	34.18	27.87	4.50	0.00
Typical sandy gley soils	**DC**	50.13	25.61	20.88	3.38	0.00
Typical sandy gley soils	**FA**	14.13	43.31	23.54	19.03	0.00
Typical sandy gley soils	**GC**	35.13	35.90	19.51	9.46	0.00
Typical sandy gley soils	**HC**	32.09	32.79	17.83	17.29	0.00
Typical sandy gley soils	**LE**	5.06	31.03	39.36	24.54	0.00
Typical sandy gley soils	**OR**	35.13	35.90	19.51	9.46	0.00
Typical sandy gley soils	**OT**	48.97	37.53	10.20	3.30	0.00
Typical sandy gley soils	**PG**	38.49	39.33	17.00	5.18	0.00

Typical sandy gley soils	**RC**	29.39	30.04	32.65	7.92	0.00
Typical sandy gley soils	**RG**	61.90	21.08	11.46	5.56	0.00
Typical sandy gley soils	**T?**	18.36	56.29	20.40	4.95	0.00
Typical sandy gley soils	**UG**	41.08	41.98	11.41	5.53	0.00
Typical stagnogley soils	**AR**	0.00	2.00	10.87	22.73	64.41
Typical stagnogley soils	**CO**	0.00	9.81	39.98	19.39	30.83
Typical stagnogley soils	**DC**	0.00	9.81	39.98	19.39	30.83
Typical stagnogley soils	**FA**	0.00	4.81	13.08	31.71	50.41
Typical stagnogley soils	**GC**	0.00	4.94	13.42	19.53	62.11
Typical stagnogley soils	**HC**	0.00	5.58	15.17	44.15	35.10
Typical stagnogley soils	**LE**	0.00	2.56	16.21	30.33	50.90
Typical stagnogley soils	**OR**	0.00	3.77	10.24	14.90	71.08
Typical stagnogley soils	**OT**	0.00	21.16	28.76	27.90	22.18
Typical stagnogley soils	**PG**	0.00	16.68	36.05	32.97	14.30
Typical stagnogley soils	**RC**	0.00	3.42	18.58	13.52	64.48
Typical stagnogley soils	**RG**	0.00	7.16	19.47	28.33	45.04
Typical stagnogley soils	**T?**	0.00	16.06	29.10	21.17	33.66
Typical stagnogley soils	**UG**	0.00	13.37	18.17	26.43	42.03

Appendix C - Chapter 4

C.1 Heuristics and Biases

There are a number of issues, known as heuristics and biases, which can negatively affect the results of the elicitation process (Tversky & Kahneman, 1974). By being aware of these potential pitfalls, and making the experts aware, it is possible to minimise the biases which can cloud expert judgement. Bias is introduced when an expert allows irrelevant information to inform decision making or fails to include take relevant information into account. There are two forms of bias; motivational and cognitive (Skinner, 1999). Motivational bias can be ascribed to the circumstances of the expert, for instance overconfidence can often be due to the expert wanting to appear certain in their opinion, hence underlining their expertise (Renooij, 2001). Also, it is much more difficult to judge extreme events as opposed to the median or mean. Experts tend to be overconfident towards the upper and lower probability ranges. Cognitive bias is attributed to heuristics, which are the intellectual 'rules of thumb' or 'mental shortcuts' that people rely on for decision making. There is an enormous amount of literature concerning heuristics and biases of the elicitation process (Kynn, 2008). This study highlights some of the commonly occurring cognitive biases, which may be particularly applicable to digital soil mapping applications. Renooij (2001) identifies four categories of bias.

Availability: Common events are brought to mind more readily than infrequent ones. This is not usually problematic, as more common events generally have a higher probability of occurring. However, memorable (but unusual/infrequent) events can skew this perception. (For instance, news of a volcanic eruption can make people believe that the event is significantly more probably than it is in reality).

Anchoring and adjustment: Experts will often provide their opinion on a range of variables by choosing an initial value of one variable or scenario and altering predictions accordingly making values bias towards the starting point. The term 'anchor' refers to the fact that usually, the expert will not adjust their estimates sufficiently (Slovic, 1972). Garthwaite et al. (2005) suggest that anchoring in the root cause of another heuristic bias, known as *conservatism*. This manifests itself as a lack of adjustment given the unconditional probability of event B in light of evidence of event A. In these terms, this is the opposite of the judgement by representativeness fallacy.

Representativeness: This heuristic is often employed when assessing a conditional probability, such as the probability of event B given event A, in statistical notation $P(B|A)$. Frequently, an expert making this judgement will compare events A and B and decide on a probability based on the likelihood of A causing B. The problem with this is that it does not necessarily account for the probability of B $P(B)$. If there is a strong correlation between A and B, it is likely that the expert will decide that event B is more probable than it would be if they were to consider the unconditional probability of event B. To clarify, if B is a rare event, then even strong evidence to suggest it may have occurred (event A) should not detract from the fact that it is unlikely. Overestimating probability in this manner is known as base rate neglect (Garthwaite et al., 2005). In terms of the practical application of probability theory, this related to the *conjunction fallacy* which states that the joint probability of event P(A,B) can never be more than the probability of the either event P(A) or P(B). Judgement by representativeness also encompasses *insensitivity to sample size*, a difficult concept to work into elicitation. Experts should be aware that the outcome of dealing with a small sample will deviate from the mean more frequently than a larger sample. Likewise, *the law of small*

numbers is the incorrect assumption that findings from a small study (in terms of area) or small number of samples will translate to a larger number (and vice versa). Conservatism and the law of small numbers suggest that people place too much weight on empirical evidence as a representation of probability distribution (Garthwaite et al., 2005).

Control: People tend to believe that they can affect an event, when actually they have no power over it at all, in turn this can lead to overconfidence (Renooij, 2001).

C.2 Predicting soil bulk density in Ireland Questionnaire

A copy of the questionnaire used to elicit conditional probabilities is attached below.

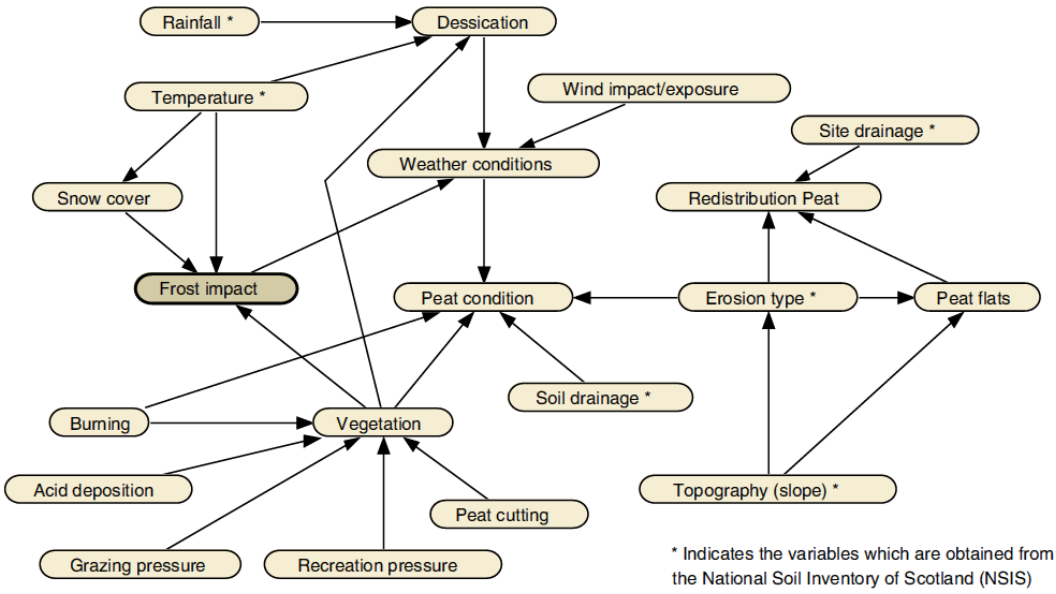
Predicting Soil Bulk Density in Ireland Questionnaire

Soil bulk density is defined as the oven-dry mass per unit volume of a soil, usually measured in g cm^{-3}. To give some context for soil surveyors, soils with the highest bulk density would typically have high sand content (70% and over), with a significant amount of clay, whilst those highest in organic matter (peat) would tend to have the lowest bulk density.

We are specifically interested in topsoil bulk density.

Your opinions will be used to predict soil bulk density at the landscape scale, meaning we are trying to identify large-scale, general trends. Please take this into consideration when making your judgements. After each question, there is a comments box for any additional information which you feel may be relevant.

The first task is to develop a conceptual diagram which represents the relationships between landscape variables and soil properties.



The Figure 1 (above) shows an example conceptual diagram developed to assess the risk of peat erosion

The purpose of the conceptual diagram is to develop an idea of how a range of landscape variables are linked in the landscape. Each variable can either be linked to one or many other variables.

The focus of this exercise is not ‘how to predict soil bulk density’ but rather to place bulk density in a landscape context.

This is far from a complete picture of processes in the landscape; it is hugely simplified as the variables chosen were limited by data availability. However, using statistical

analysis, all the variables used in Figure 3 have been shown to either directly or indirectly influence soil bulk density.

Please note that feedback loops are not allowed (Figure 2)

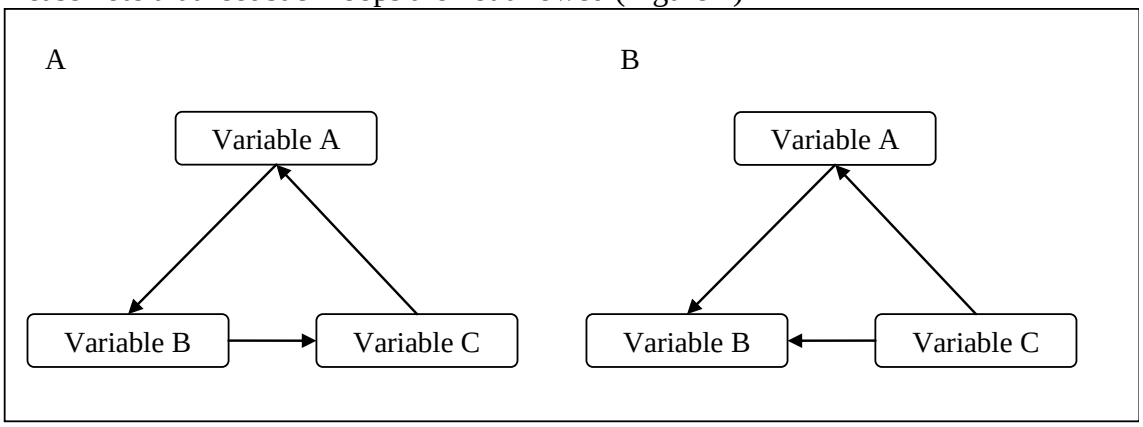
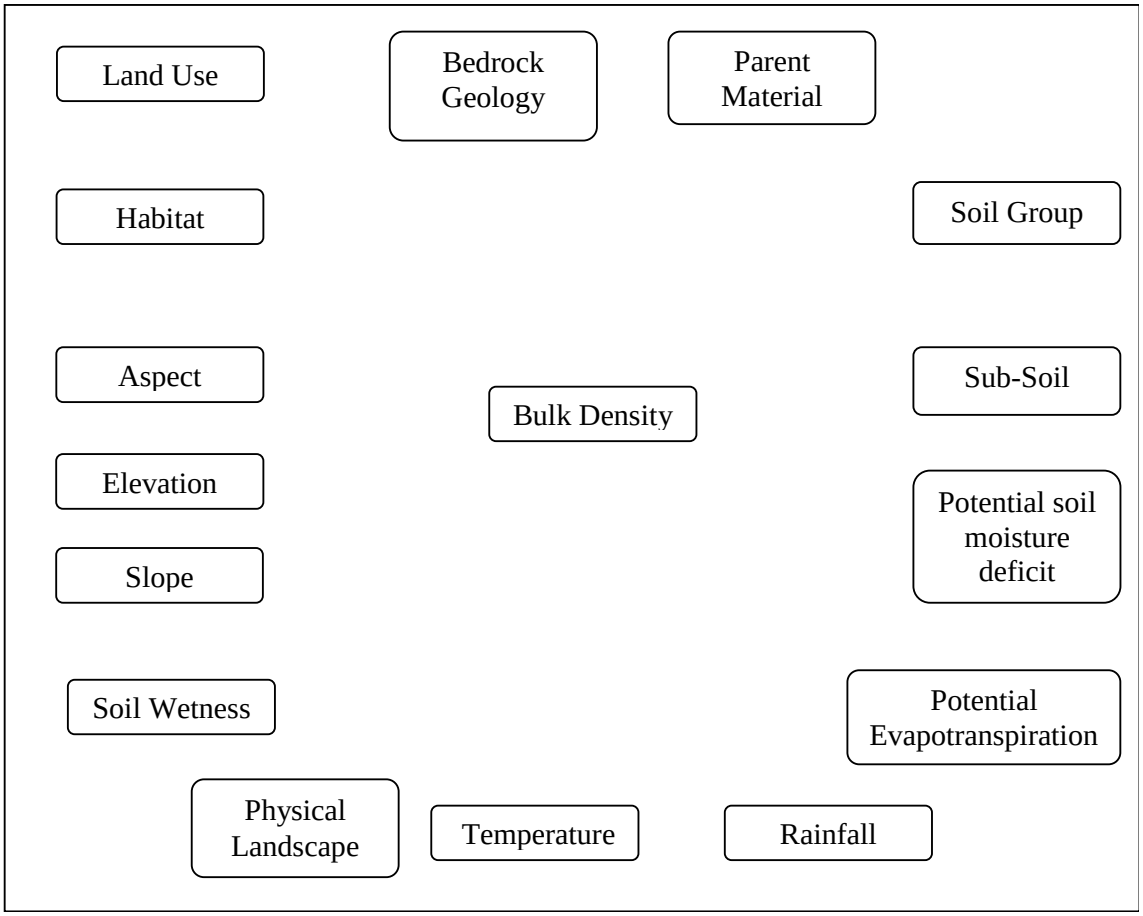


Figure 2: A) is an example of a feedback loop which is not allowed within the conceptual model. B) is an example of a valid relationship within the conceptual model

Conceptual diagram



Please complete the conceptual diagram by drawing arrows to indicate the relationship between variables. The direction of the arrow indicates the direction

of influence. Not all variables need to link directly to bulk density and each variable can link to as many or few variables as you see fit (while still following the structural rules in Figure2).

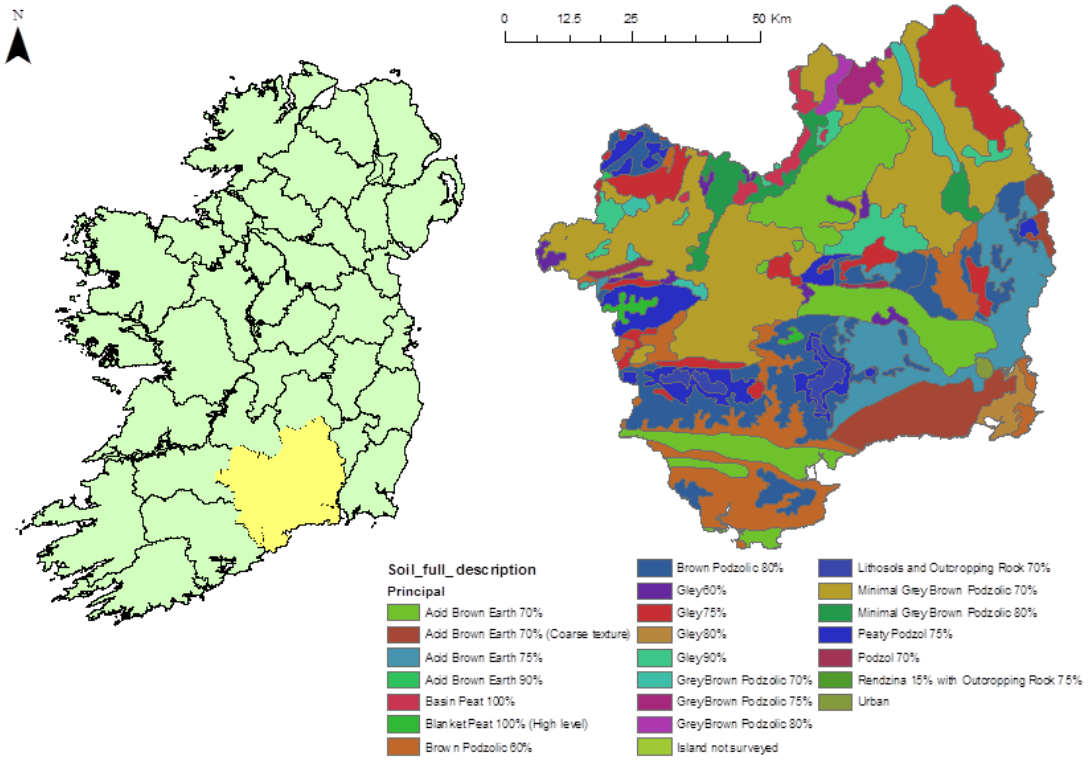


Figure 3: Study area. The study area comprises the three counties of South Tipperary, Waterford and Kilkenny. The distribution of the principal soils is shown on the right of the figure.

Background Information

Professional/Academic background (e.g. soil scientist/ soil surveyor/ geologist)		
Experience: Length of time you have worked with soils?		
Do you have experience of working with soils outside of Ireland? If 'yes', please specify location and duration	Yes No	
How long have you worked with Irish soils?		
Do you have fieldwork experience in South Tipperary/Waterford/ Kilkenny? If 'yes', please specify duration	Yes No	
Before this study, were you familiar with soil bulk density?	Yes No	

Bulk Density

For use in a Bayesian Network, it is necessary to separate continuous variables into a number of classes. The table below shows the class boundaries for soil bulk density. As you may not be familiar with the numerical values, we have provided some 'example soils' which (on average) will have bulk density values within the specified ranges.

Note that this is for the UK (the data is not available for Ireland at present)

Class	Value (g cm^{-3})	Example soil (England & Wales Classification)
Very high	Over 1.5	Brown sands
High	1.2-1.5	Argillic brown earths
Medium	0.9-1.2	Typical stagnogley soils
Low	0.6-0.9	Stagnohumic gley soils
Very Low	Less than 0.6	Peats

As experts, you are all familiar with the soils of Kilkenny, Waterford and South Tipperary (Figure 3). For each of the following scenarios, please imagine you are talking 100 samples from across the study area. For each state (e.g. GSM1, GSM3 etc in the 'soil' variable below), please estimate the number of samples which would fall into each bulk density class.

Each horizontal row should sum to 100

Remember, there are no 'right answers', the only aim is for us to represent your understanding given the limited information you are provided with.

Soil	Very High	High	Medium	Low	Very Low
GSM1 Peaty Podzol 75% Lithosol (15%) Blanket Peat (10%)					
GSM3 Lithosols and Outcropping Rock 70% Blanket Peat (25%) Peaty Podzol (5%)					
GSM5 Brown Podzolic 80% Gley (15%) Podzol (5%)					
GSM8 Brown Podzolic 80% Gley (15%) Podzol (5%)					
GSM11 Acid Brown Earth 70% (Coarse texture) Gley (25%) Podzol (5%)					
GSM12 Acid Brown Earth 70% Grey Brown Podzolic (15%) Gley (15%)					
GSM13 Acid Brown Earth 75% Gley (15%) Brown Podzolic (10%)					
GSM14 Brown Podzolic 60% Acid Brown Earth (20%) Gley (20%)					
GSM18 Acid Brown Earth 70% Gley (5%) Peaty Gley (5%)					
GSM20 Gley 75% Peaty Gley (25%)					
GSM21 Gley 75% Acid Brown Earth (15%) Peat (10%)					
GSM28 Grey Brown Podzolic 70% Brown Earth (20%) Gley (5%) Basin Peat (5%)					
GSM29 Minimal Grey Brown Podzolic 80% Gley (10%) Brown Earth (5%) Basin Peat (5%)					
GSM32 Minimal Grey Brown Podzolic 70% Gley (20%) Brown Earth (10%)					
GSM37 Gley 90% Grey Brown Podzolic (10%)					
GSM38 Gley 80% Grey Brown Podzolic (20%)					
GSM40 Gley 60% Brown Earth (20%) Peaty Gley (20%)					
GSM41 Basin Peat 100%					

Land Use	Very High	High	Medium	Low	Very Low
COR7 Complex cultivation patterns*					
COR8 Coniferous forests					
COR18 Land principally occupied by agriculture with significant areas of natural vegetation					
COR23 Non-irrigated arable land					
COR24 Pastures					
COR25 Peat bogs					
COR32 Transitional woodland scrub					

* Small parcels of diverse annual crops, pasture and/or permanent crops. This can include arable land, pasture, orchards and city gardens (each under 25 ha)

Bedrock Geology	Very High	High	Medium	Low	Very Low
GEO1 Igneous					
GEO3 Igneous granites					
GEO4 Limestone					
GEO5 Metamorphic					
GEO6 Metamorphic and igneous					
GEO7 Sandstone					
GEO8 Sandstone and Shales					
GEO10 Sandstones					
GEO11 Shales					

Subsoil	Very High	High	Medium	Low	Very Low
SBS2 Alluvium Fluvial					
SBS5 Anthropogenic man made					
SBS9 Bedrock limestone					
SBS10 Bedrock sandstone					
SBS11 Bedrock sandstone and shales					
SBS12 Bedrock shale slate					
SBS14 Drift igneous and metamorphic stones					
SBS15 Drift limestone					
SBS16 Drift siliceous stones					
SBS18 Peat Blanket and raised bog					
SBS19 Peat cutover and industrial					

Landscape	Very High	High	Medium	Low	Very Low
Flat to Undulating Lowland (mainly dry mineral soils)					
Flat to Undulating Lowland (mainly wet mineral and organic soils)					
Hill					
Mountain and Hill					
Rolling lowland					
Urban					

Habitat	Very High	High	Medium	Low	Very Low
BL Built Land					
CR Rocky Complex					
GAGS Dry Grassland					
GSW Wet Grassland					
H Heath					
RBF Raised Bog/Fen					
WNWD Mature Forest					
WSWL Forest (U) and scrub					

Parent Material	Very High	High	Medium	Low	Very Low
Alluvium					
Limestone glacial till					
Limestone Moranic gravels and sands					
Mixed sandstone and limestone glacial till					
Mostly granite or rhyolite glacial till					
Mostly granite sandstone					
Mostly sandstone					
Mostly sandstone granite quartzite or mica					
Ordovician Silurian Cambrian shale glacial till					
Ordovician Silurian Cambrian shales and mica schist					
Peat					
Sandstone glacial till					
Sandstone Lower Avonian shale glacial till					
Till of Irish Sea origin with limestone and shale					
Upper Carboniferous shale and Sandstone glacial till					

Upper Carboniferous shale glacial till					

For the topographic and climatic variables, you have been provided with the range of values from within the study area, plus the mean value. You may wish to indicate a cut off point for the groups, beyond which you feel there would be a noticeable effect on soil type and subsequently bulk density. This could before a single category or all three.

If you do not indicate a range, the 'high', 'medium' and 'low' groups will be split into three equal groups across the range.

Range: 0-59.68° average 3.87°

Slope	Very High	High	Medium	Low	Very Low
No Slope					
Medium slope					
Steep					

Range 0-903 m average 127.51m

Elevation	Very High	High	Medium	Low	Very Low
Low					
Medium					
High					

Aspect	Very High	High	Medium	Low	Very Low
North					
South					
East					
West					

Soil wetness index 7.15-25.92 average 15.34

Soil Wetness	Very High	High	Medium	Low	Very Low
Dry					
Average					
Wet					

Annual mean per year (mm per year) 842 -2,362 average 1,112

Rainfall	Very High	High	Medium	Low	Very Low
Low					
Medium					
High					

Annual mean temperature 4.6°C -10.46°C average 9.2°C

Temperature	Very High	High	Medium	Low	Very Low
Low					
Medium					
High					

Range 0-129.87 average 48.66

Potential soil moisture deficit	Very High	High	Medium	Low	Very Low
Low					
Medium					
High					

Range 369.8-563.8 average 506.4

Potential evapotranspiration	Very High	High	Medium	Low	Very Low
Low					
Medium					
High					

Comments:

C.3 Hierarchical model input variable reclassification

For use in the Hierarchical model, the landscape variables soil group, subsoil, parent material, land use, habitat and Physiographic division were reclassified in to 4 classes: High, Medium, Low, Very Low. This reclassification was carried out using the consensus of expert opinion. The reclassified variables are shown in Table C.3-1.

Table C.3-1: Reclassified variables for use in the Hierarchical Bayesian Network

	High	Medium	Low	Very Low
Soil group	GSM12, GSM13, GSM32, GSM37, GSM38, GSM40, GSM43	GSM11, GSM14, GSM18, GSM20, GSM21, GSM28, GSM29	GSM5	GSM1, GSM3, GSM 41
Subsoil	SBS5, SBS9	SBS2, SBS10, SBS11, SBS12, SBS15, SBS16	SBS14	SBS18, SBS19
Parent Material	Alluvium, Limestone glacial till, Mixed sandstone and limestone glacial till	Limestone moranic gravels and sands, Mostly granite or rhyolite glacial till, Mostly sandstone, Mostly sandstone granite quartzite or mica, Ordovician Silurian Cambrian shale glacial till, Sandstone glacial till, Sandstone Lower Avonian shale glacial till, Till of Irish Sea origin with limestone and shale, Upper Carboniferous shale and Sandstone glacial till, Upper Carboniferous shale glacial till, Sandstone glacial till	Ordovician Silurian Cambrian shales and mica schist	Peat
Corine		COR7, COR18, COR23, COR24	COR32	COR8, COR25
Habitat	BL	GAGS, GSW,	CR, WNWD	H, RBF, WSWL
Physio		Flat to Undulating (Dry), Flat to Undulating (Wet), Rolling Lowland	Hill	Mountain and Hill

C.4 Naive Bayesian Network CPTs

The expert-derived conditional probability tables for the naive model are shown below. The CPTs in the Naive BN were derived from a mathematical mean of individual opinions.

Table C.4-1: CPT for the 'GSM' node of the Naive BN

Soil	Very High	High	Medium	Low	Very Low
GSM1	0	7	1	40	52
GSM3	9	25	10.4	34	21.6
GSM5	8	36	42	12	2
GSM11	28	16	40	12	4
GSM12	29	16	43	10	2
GSM13	30	15	39	14	2
GSM14	7	26	47	18	2
GSM18	18	16	36	22	8
GSM20	15	28	35	12	10
GSM21	15	28	41	7	9
GSM28	14	35	35	9	7
GSM29	10	23	55	5	7
GSM32	13	41	43	3	0
GSM37	15	46	39	0	0
GSM38	13	41	45	1	0
GSM40	12	34	36	16	2
GSM41	0	0	0	18	82

Table C.4-2: CPT for the 'Corine' land use node of the Naive BN

Land_use	Very High	High	Medium	Low	Very Low
COR7	12.5	35	37.5	10	5
COR8	7	18	20	35	20
COR18	5	26.875	42.5	15.625	10
COR23	5	25	42.5	18.75	8.75
COR24	1	24	49	20	6
COR25	0	0	8	22	70
COR32	7.5	26.25	37.5	21.25	7.5

Table C.4-3: CPT for the 'GEO' bedrock Geology node of the Naive BN

Bedrock_Geology	Very High	High	Medium	Low	Very Low
GEO1	7.5	26.25	50	15	1.25
GEO3	16	34	33	14	3
GEO4	1	32	50	14	3
GEO5	7	25	48	18	2
GEO6	15	23	42	18	2
GEO7	18	24	33	18	7
GEO8	13	32	34.5	15.5	5
GEO11	6	30	38	24	2

Table C.4-4: CPT for the 'Subsoil' node of the Naive BN

Subsoil	Very High	High	Medium	Low	Very Low
SBS2	20	34	32	9	5
SBS5	23	50	18	5	4
SBS9	13.75	33.75	41.25	8.75	2.5
SBS10	18	28	44	7	3
SBS11	18	33	33	12	4
SBS12	24	29.5	34	11.5	1
SBS14	14	27	34	24	1
SBS15	17.5	28.5	33	19	2
SBS16	10	27.5	46.25	15	1.25
SBS18	0.5	0.5	2.5	26.5	70
SBS19	0.5	0.5	4.5	28.5	66

Table C.4-5: CPT for the 'Physio' physiographic landscape unit node of the Naive BN

Landscape	Very High	High	Medium	Low	Very Low
Flat to Undulating Lowland (mainly dry mineral soils)	10	37	34	13	6
Flat to Undulating Lowland (mainly wet mineral and organic soils)	11	32	27	16	14
Hill	3	18	43	30	6
Mountain and Hill	3	22	36	20	19
Rolling lowland	6.67	46.67	33.33	10	3.33
Urban	21.67	30	36.67	10	1.67

Table C.4-6: CPT for the 'Habitat' node of the Naive BN

Habitat	Very High	High	Medium	Low	Very Low
BL Built Land	25	30	30	13.33	1.67
CR Rocky Complex	7.5	20	35	30	7.5
GAGS Dry Grassland	10	26	52	10	2
GSW Wet Grassland	12	28	33	24	3
H Heath	6.67	6.67	16.67	33.33	36.67
RBF Raised Bog/Fen	0	4	8	30	58
WNWD Mature Forest	2	17	32	26	23
WSWL Forest (U) and scrub	6.25	10	32.5	27.5	23.75

Table C.4-7: CPT for the 'Parent Material' node of the Naive BN

Parent Material	Very High	High	Medium	Low	Very Low
Alluvium	17	32	32	13	6
Limestone glacial till	6	28	47	15	4
Limestone Moranic gravels and sands	19	17	43	17	4
Mixed sandstone and limestone glacial till	11	26	45	12	6
Mostly granite or rhyolite glacial till	11	31	37	14	7
Mostly sandstone	14	26	46	10	4
Mostly sandstone granite quartzite or mica	4	23	53	15	5
Ordovician Silurian Cambrian shale glacial till	10	26	47	12	5
Ordovician Silurian Cambrian shales and mica schist	7.5	32.5	45	10	5
Peat	0	0	2	24	74
Sandstone glacial till	10	27	40	16	7
Sandstone Lower Avonian shale glacial till	5	30	42.5	15	7.5
Till of Irish Sea origin with limestone and shale	10	32	42	12	4
Upper Carboniferous shale and Sandstone glacial till	5	23	43	21	8
Upper Carboniferous shale glacial till	6.25	30	32.5	20	11.25

Table C.4-8: CPT for the 'Slope' node of the Naive BN

Slope	Very High	High	Medium	Low	Very Low
No Slope	5	20	25	20	30
Medium slope	7	19	31	28	15
Steep	2.5	15	38.75	33.75	10

Table C.4-9: CPT for the 'Elevation' node of the Naive BN

Elevation	Very High	High	Medium	Low	Very Low
Low	6	23	30	24	17
Medium	1	24	45	24	6
High	1	12	31	26	30

Table C.4-10: CPT for the 'Aspect' node of the Naive BN

Aspect	Very High	High	Medium	Low	Very Low
North	15	35	25	15	10
South	0	15	50	20	15
East	0	30	40	30	0
West	0	30	40	30	0

Table C.4-11: CPT for the 'SWI' Soil Wetness Index node of the Naive BN

Soil Wetness	Very High	High	Medium	Low	Very Low
Dry	5	10	52.5	20	12.5
Average	0	27.5	50	17.5	5
Wet	15	32.5	30	12.5	10

Table C.4-12: CPT for the 'Rainfall' node of the Naive BN

Rainfall	Very High	High	Medium	Low	Very Low
Low	3.75	18.75	48.75	15	13.75
Medium	3.75	22.5	46.25	23.75	3.75
High	13.33333	33.33333	28.33333	10	15

Table C.4-13: CPT for the 'Temperature' node of the Naive BN

Temperature	Very High	High	Medium	Low	Very Low
Low	5	20	25	25	25
Medium	2.5	22.5	42.5	22.5	10
High	2.5	27.5	42.5	17.5	10

Table C.4-14: CPT for the 'PSMD' potential soil moisture deficit node of the Naive BN

Potential Soil Moisture Deficit	Very High	High	Medium	Low	Very Low
Low	23.75	31.25	23.75	13.75	7.5
Medium	5	31.67	46.67	13.33	3.33
High	0	20	30	20	30

Table C.4-15: CPT for the 'PT' potential evapotranspiration node of the Naive BN

Potential evapotranspiration	Very High	High	Medium	Low	Very Low
Low	10	43.33	36.67	10	0
Medium	1.67	30	50	16.67	1.67
High	0	20	36.67	20	23.33

C.5 Naive Bayesian Network Structure

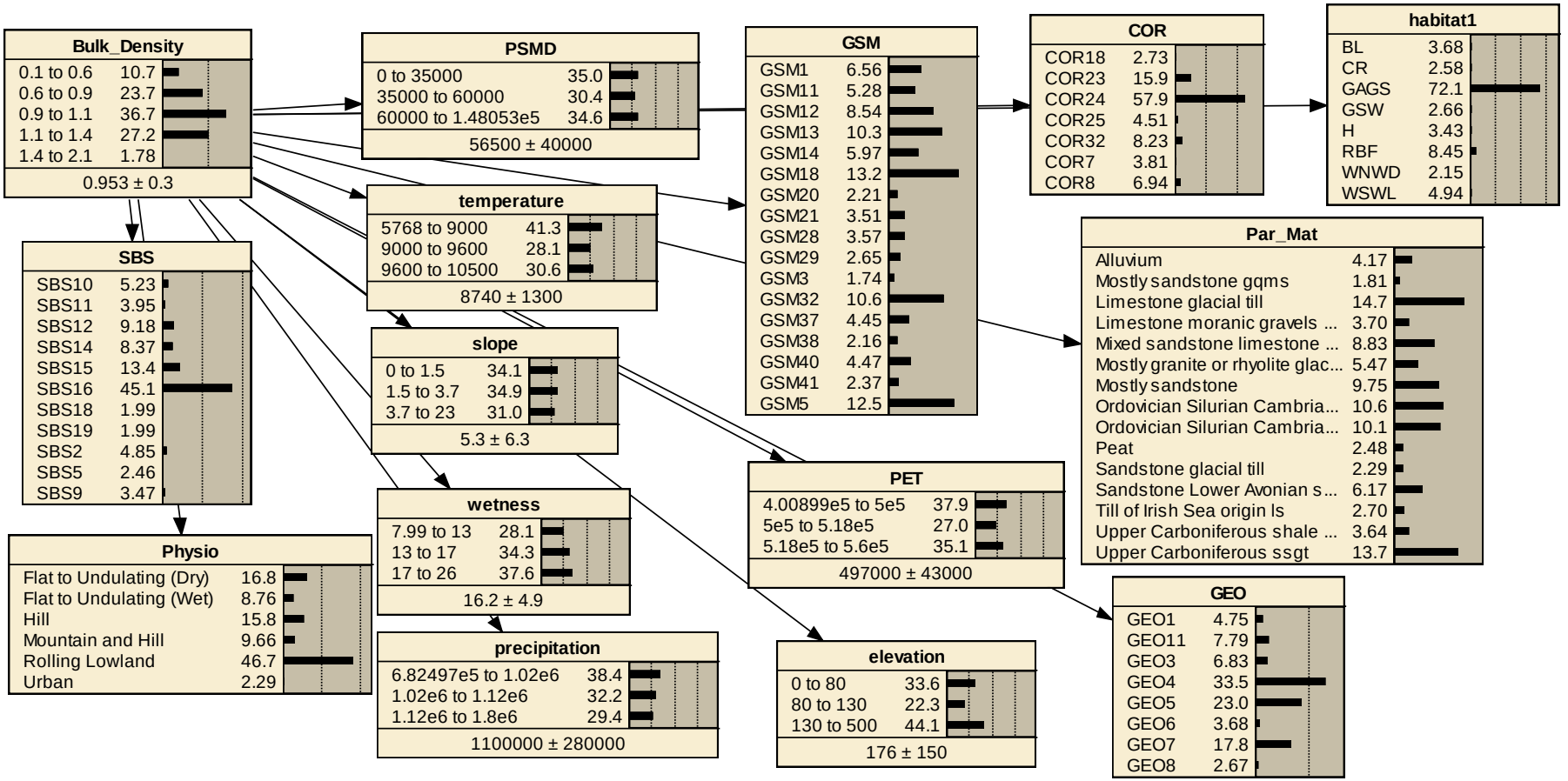


Figure C.5-1: The Naive Network with Expert Derived CPTs

C.6 Hierarchical Bayesian Network CPTs

The expert-derived conditional probability tables for the hierarchical model are shown below. The CPTs in the Hierarchical BN were derived from a group consensus.

Table C.6-1: CPT for the 'Bulk_Density' node of the Hierarchical BN

Soil	Land Use	Climate	Very Low	Low	Medium	High
Very Low	**Very Low**	**Low**	100	0	0	0
Very Low	**Very Low**	**Medium**	100	0	0	0
Very Low	**Very Low**	**High**	100	0	0	0
Very Low	**Low**	**Low**	90	10	0	0
Very Low	**Low**	**Medium**	90	10	0	0
Very Low	**Low**	**High**	90	10	0	0
Very Low	**Medium**	**Low**	90	10	0	0
Very Low	**Medium**	**Medium**	90	10	0	0
Very Low	**Medium**	**High**	90	10	0	0
Very Low	**High**	**Low**	90	10	0	0
Very Low	**High**	**Medium**	90	10	0	0
Very Low	**High**	**High**	90	10	0	0
Low	**Very Low**	**Low**	95	5	0	0
Low	**Very Low**	**Medium**	95	5	0	0
Low	**Very Low**	**High**	95	5	0	0
Low	**Low**	**Low**	20	50	30	0
Low	**Low**	**Medium**	20	50	30	0
Low	**Low**	**High**	20	50	30	0
Low	**Medium**	**Low**	10	50	40	0
Low	**Medium**	**Medium**	10	50	40	0
Low	**Medium**	**High**	10	50	40	0
Low	**High**	**Low**	10	50	40	0
Low	**High**	**Medium**	10	50	40	0
Low	**High**	**High**	10	50	40	0
Medium	**Very Low**	**Low**	90	10	0	0
Medium	**Very Low**	**Medium**	90	10	0	0
Medium	**Very Low**	**High**	90	10	0	0
Medium	**Low**	**Low**	10	30	50	10
Medium	**Low**	**Medium**	5	25	50	20
Medium	**Low**	**High**	5	25	50	20
Medium	**Medium**	**Low**	0	15	60	25
Medium	**Medium**	**Medium**	0	15	60	25
Medium	**Medium**	**High**	0	15	60	25
Medium	**High**	**Low**	0	15	60	25
Medium	**High**	**Medium**	0	15	60	25

Medium	**High**	**High**	0	15	60	25
High	**Very Low**	**Low**	80	15	5	0
High	**Very Low**	**Medium**	80	15	5	0
High	**Very Low**	**High**	80	15	5	0
High	**Low**	**Low**	0	10	40	50
High	**Low**	**Medium**	0	5	40	55
High	**Low**	**High**	0	5	35	60
High	**Medium**	**Low**	0	10	30	60
High	**Medium**	**Medium**	0	0	30	70
High	**Medium**	**High**	0	0	30	70
High	**High**	**Low**	0	0	30	70
High	**High**	**Medium**	0	0	25	75
High	**High**	**High**	0	0	20	80

Table C.6-2: CPT for the 'Soil' node of the Hierarchical BN

GSM	Subsoil	Parent_Material	Very_Low	Low	Medium	High
Very Low	**Very Low**	**Very Low**	100	0	0	0
Very Low	**Very Low**	**Low**	100	0	0	0
Very Low	**Very Low**	**Medium**	100	0	0	0
Very Low	**Very Low**	**High**	100	0	0	0
Very Low	**Low**	**Very Low**	100	0	0	0
Very Low	**Low**	**Low**	100	0	0	0
Very Low	**Low**	**Medium**	100	0	0	0
Very Low	**Low**	**High**	95	5	0	0
Very Low	**Medium**	**Very Low**	95	5	0	0
Very Low	**Medium**	**Low**	95	5	0	0
Very Low	**Medium**	**Medium**	95	5	0	0
Very Low	**Medium**	**High**	95	5	0	0
Very Low	**High**	**Very Low**	90	10	0	0
Very Low	**High**	**Low**	90	10	0	0
Very Low	**High**	**Medium**	85	10	5	0
Very Low	**High**	**High**	85	10	5	0
Low	**Very Low**	**Very Low**	20	75	5	0
Low	**Very Low**	**Low**	20	75	5	0
Low	**Very Low**	**Medium**	20	75	5	0
Low	**Very Low**	**High**	20	75	5	0
Low	**Low**	**Very Low**	15	75	10	0
Low	**Low**	**Low**	15	75	10	0
Low	**Low**	**Medium**	15	75	10	0
Low	**Low**	**High**	15	75	10	0
Low	**Medium**	**Very Low**	15	70	15	0
Low	**Medium**	**Low**	15	70	15	0

Low	**Medium**	**Medium**	15	70	15	0
Low	**Medium**	**High**	15	70	15	0
Low	**High**	**Very Low**	10	65	25	0
Low	**High**	**Low**	10	65	25	0
Low	**High**	**Medium**	10	65	25	0
Low	**High**	**High**	10	65	25	0
Medium	**Very Low**	**Very Low**	5	30	60	5
Medium	**Very Low**	**Low**	5	30	60	5
Medium	**Very Low**	**Medium**	5	30	60	5
Medium	**Very Low**	**High**	5	30	60	5
Medium	**Low**	**Very Low**	5	20	70	5
Medium	**Low**	**Low**	5	20	70	5
Medium	**Low**	**Medium**	5	20	70	5
Medium	**Low**	**High**	5	20	70	5
Medium	**Medium**	**Very Low**	5	10	70	15
Medium	**Medium**	**Low**	5	10	70	15
Medium	**Medium**	**Medium**	5	10	70	15
Medium	**Medium**	**High**	5	10	70	15
Medium	**High**	**Very Low**	0	10	70	20
Medium	**High**	**Low**	0	10	70	20
Medium	**High**	**Medium**	0	10	70	20
Medium	**High**	**High**	0	10	70	20
High	**Very Low**	**Very Low**	0	10	55	35
High	**Very Low**	**Low**	0	10	55	35
High	**Very Low**	**Medium**	0	10	55	35
High	**Very Low**	**High**	0	10	55	35
High	**Low**	**Very Low**	0	5	60	35
High	**Low**	**Low**	0	5	60	35
High	**Low**	**Medium**	0	5	60	35
High	**Low**	**High**	0	5	60	35
High	**Medium**	**Very Low**	0	0	40	60
High	**Medium**	**Low**	0	0	40	60
High	**Medium**	**Medium**	0	0	40	60
High	**Medium**	**High**	0	0	40	60
High	**High**	**Very Low**	0	0	30	70
High	**High**	**Low**	0	0	30	70
High	**High**	**Medium**	0	0	30	70
High	**High**	**High**	0	0	30	70

Table C.6-3: CPT for the 'Land_Use' node of the Hierarchical BN

Physio	Corine	Habitat	Very_Low	Low	Medium	High
Very Low	**Very Low**	**Very Low**	100	0	0	0
Very Low	**Very Low**	**Low**	100	0	0	0
Very Low	**Very Low**	**Medium**	100	0	0	0
Very Low	**Very Low**	**High**	100	0	0	0
Very Low	**Low**	**Very Low**	100	0	0	0
Very Low	**Low**	**Low**	100	0	0	0
Very Low	**Low**	**Medium**	100	0	0	0
Very Low	**Low**	**High**	100	0	0	0
Very Low	**Medium**	**Very Low**	95	5	0	0
Very Low	**Medium**	**Low**	90	10	0	0
Very Low	**Medium**	**Medium**	90	10	0	0
Very Low	**Medium**	**High**	90	10	0	0
Very Low	**High**	**Very Low**	95	5	0	0
Very Low	**High**	**Low**	90	10	0	0
Very Low	**High**	**Medium**	85	10	5	0
Very Low	**High**	**High**	85	10	5	0
Low	**Very Low**	**Very Low**	95	5	0	0
Low	**Very Low**	**Low**	90	10	0	0
Low	**Very Low**	**Medium**	90	10	0	0
Low	**Very Low**	**High**	90	10	0	0
Low	**Low**	**Very Low**	95	5	0	0
Low	**Low**	**Low**	20	60	20	0
Low	**Low**	**Medium**	20	60	20	0
Low	**Low**	**High**	20	60	20	0
Low	**Medium**	**Very Low**	20	60	20	0
Low	**Medium**	**Low**	20	60	20	0
Low	**Medium**	**Medium**	5	30	50	15
Low	**Medium**	**High**	5	30	50	15
Low	**High**	**Very Low**	5	30	50	15
Low	**High**	**Low**	5	30	50	15
Low	**High**	**Medium**	5	30	50	15
Low	**High**	**High**	5	30	50	15
Medium	**Very Low**	**Very Low**	100	0	0	0
Medium	**Very Low**	**Low**	95	5	0	0
Medium	**Very Low**	**Medium**	95	5	0	0
Medium	**Very Low**	**High**	95	5	0	0
Medium	**Low**	**Very Low**	90	10	0	0
Medium	**Low**	**Low**	20	50	30	0
Medium	**Low**	**Medium**	20	50	30	0
Medium	**Low**	**High**	20	50	30	0
Medium	**Medium**	**Very Low**	80	10	10	0

Medium	**Medium**	**Low**	0	20	60	20
Medium	**Medium**	**Medium**	0	20	60	20
Medium	**Medium**	**High**	0	20	60	20
Medium	**High**	**Very Low**	80	10	10	0
Medium	**High**	**Low**	0	10	60	30
Medium	**High**	**Medium**	0	10	60	30
Medium	**High**	**High**	0	10	60	30
High	**Very Low**	**Very Low**	95	5	0	0
High	**Very Low**	**Low**	95	5	0	0
High	**Very Low**	**Medium**	90	10	0	0
High	**Very Low**	**High**	90	10	0	0
High	**Low**	**Very Low**	90	10	0	0
High	**Low**	**Low**	10	70	20	0
High	**Low**	**Medium**	0	10	70	20
High	**Low**	**High**	0	10	70	20
High	**Medium**	**Very Low**	85	10	5	0
High	**Medium**	**Low**	0	20	50	30
High	**Medium**	**Medium**	0	10	40	50
High	**Medium**	**High**	0	0	40	60
High	**High**	**Very Low**	80	15	5	0
High	**High**	**Low**	0	25	50	25
High	**High**	**Medium**	0	0	40	60
High	**High**	**High**	0	0	40	60

Table C.6-4: CPT for the 'Climate' node of the Hierarchical BN

PSMD	PET	Wetness	Low	Medium	High
Low	Low	Dry	10	60	30
Low	Low	Average	10	70	20
Low	Low	Wet	50	30	20
Low	Medium	Dry	15	70	15
Low	Medium	Average	20	60	20
Low	Medium	Wet	50	30	20
Low	High	Dry	10	60	30
Low	High	Average	15	70	15
Low	High	Wet	60	30	10
Medium	Low	Dry	10	60	30
Medium	Low	Average	15	70	15
Medium	Low	Wet	30	50	20
Medium	Medium	Dry	10	60	30
Medium	Medium	Average	20	60	20
Medium	Medium	Wet	30	50	20
Medium	High	Dry	10	60	30

Medium	High	Average	20	60	20
Medium	High	Wet	30	50	20
High	Low	Dry	10	60	30
High	Low	Average	15	70	15
High	Low	Wet	50	30	20
High	Medium	Dry	10	60	30
High	Medium	Average	20	60	20
High	Medium	Wet	30	50	20
High	High	Dry	20	50	30
High	High	Average	20	50	30
High	High	Wet	60	30	10